

R语言与统计分析

汤银才 主编

高等教育出版社

二〇〇八年五月

内容介绍

本书以数据的常用统计分析方法为基础，在简明扼要地阐述统计学基本概念、基本思想与基本方法的基础上，讲述与之相对应的**R**函数的实现，并通过具体的例子说明统计问题求解的过程。

本书注重思想性、实用性和可操作性。在内容的安排上不仅包含了基础统计分析中的探索性数据分析、参数的估计与假设检验，还包括的非参数统计分析的常用方法、多元统计分析方法及贝叶斯统计分析方法。每一部分都通过具体例子重点讲述解决问题的思想、方法和在**R**中的实现过程。通过本书读者不仅可以快速学会**R**的基本原理与核心内容，而且根据提供的例子与相应的**R**程序学会解决问题的统计计算方法与基本的编程技术，为解决更为复杂的统计问题奠定扎实的基础。

本书可作为各专业本科生、研究生数理统计或应用统计课程的基础教材或实验教材，也可作为从事数据统计分析研究人员、工程技术人员的工具书或参考读物。

前言

统计学的任务是研究有关收集、整理、分析数据,从而对所考察的问题作出一定的结论的方法与理论. 作为一门科学,统计学有其坚实的理论基础,研究统计学方法的理论基础问题的那一部分,构成了所谓数理统计学的内容. 其次,统计学就其本质来讲,是一门实用性很强的科学,它在人类活动的各个领域有着广泛的应用. 因此数理统计的理论与方法应该与实际相结合,解决社会、经济、工农业生产、生物制药、航空航天、质量管理、环境资源等领域中的各种问题. 最后,统计学又是一门技术性很强的科学,由于所研究问题越来越复杂、变量之间关联性越强、数据的规模越来越大,使得原有的计算方法无法实现. 现在,随着计算机的不断发展与普及,特别是近20年来统计计算的突破性进展及统计软件的不断完善和成熟,使得解决这些问题不仅成为可能,而且越来越容易、快速.

目前许多大学几乎所有的理工科,甚至文科的许多专业都开设了《数理统计》或《应用统计》之类的课程,有的还编写了相应的教材,这是令人可喜的. 这些课程与教材的共同特点是以较大的篇幅介绍数理统计的理论、方法与实际背景,并配有一定数量的例子和习题. 部分学校还为有统计专业和应用数学专业的学生开设SAS或Matlab统计软件,为经济统计专业的学生开设SPSS或EViews统计软件,但这还远远不够.

作者长期从事概率论与数理统计、统计计算及统计软件的教学工作,我们发现目前的统计教学普遍存在的问题有: 一、关于教学内容: 在有限的课时下,对于非统计专业的学生采用统计专业学生的教学方式,过多强调理论的重要性,从而忽视了统计思想和数据处理能力的培养; 有的因为仅用一学期(54课时或更少)讲授概率论与数理统计,面面俱到的概率论教学使学生无法学到诸如回归分析与方差分析的重要内容. 二、关于软件教学: 由于没有软件支持,使用传统的教学方法和教材,无论是老师讲解例题,还是学生完成习题都要花费大量的时间进行手工计算,且错误率高. 使用软件可使数据分析更具

直观性、灵活性和可重复性,可起到举一反三的作用,提高学生的学习兴趣和动手(操作或编程)能力. 三、关于统计教学与软件教学是否分开: 统计教学与软件教学分开教学会产生一定的重复性,从而浪费有限教学课时,降低学习的效率. 分开的教学会使大部分非统计专业的学生不能得到统计软件操作和数据分析能力的培养. 有了统计软件,可大大增加教学的信息量、节省时间用于培养学生统计软件的上机操作能力;有了统计软件,使得大规模或海量数据分析和精确计算成为可能,也使教材中的许多附表(如常用分布的分位数表)失去其必要性. 四、关于R软件: 本书之所以采用R软件,主要原因是其强大的数据的图形展示和统计分析功能、免费使用和更新及大量可随时加载的有针对性的软件包. 而SAS、Matlab、SPSS、EViews却都是收费软件,与R功能几乎相同的S-PLUS也是收费的. R高效的代码、简洁的输出和强大的帮助系统使统计软件辅助的统计教学成为可能. 基于R开发的菜单式驱动的图形界面工具R Commander和PMG(见附录B)使得基础统计分析像SPSS一样容易实现.

本书介绍了R的基本功能、常用的数据处理与分析方法及它们在R中的实现. 全书共分十一章及三个附录: 第一章, R 介绍. 介绍了R软件的功能与安装. 第二章, R的基本原理与核心. 简明扼要地介绍了R软件的使用方法, 主要侧重于不同类型的数据的操作与函数的使用. 第三章, 概率与分布. 介绍了常用的离散与连续型分布及R中有关的四类函数: 分布函数、概率函数、分位数函数和随机数生存函数. 第四章, 探索性数据分析. 介绍了单组和多组数据中特征量的提取方法及数据的图形展示方法. 第五章, 参数估计. 主要介绍了单总体与两总体正态及二项分布参数的点估计与区间估计. 第六章, 参数的假设检验. 主要介绍了单总体与两总体正态及二项分布参数的假设检验. 第七章, 非参数的假设检验. 主要介绍了常用的几个非参数检验方法. 第八章, 方差分析. 主要介绍了多组数据比较的单因子与双因子方差分析及协方差分析方法. 第九章, 回归分析与相关分析. 介绍了随机变量之间关系的度量与回归分析及诊断方法. 第十章, 多元统计分析介绍. 介绍了多元分析中常用的主成分分析、因子分析、判别分析、聚类分析、典型相关分析及对应分析方法. 第十一章, 贝叶斯统计分析. 介绍了贝叶斯分析中单参数与多参数模型、分层模型及回归模型的分析方法. 最后是附录, 附录B介绍了基于R开发的基础统计分析的菜单式工具R Commander和PMG, 附录C介绍了R的3个编程环境: R WinEdt、Tinn-R及SciViews-R. 全书在所有程序都在R的2.6.0版本上调试通过, 原则上在其它版本上也可以运行.

本书的特点是: 注重统计思想、实用性和可操作性. 我们在内容的设计上尽可能简化统计理论与方法的推导过程, 对于主要的统计知识都通过一个具体

例子展开、讲清要解决问题的思想、方法和具体的实现过程. 所有方法的实现都有相应的**R**函数的调用格式, 而例子讲解的**R**程序都全部嵌入在正文中, 便于读者举一反三, 解答习题或进行其它类似的数据分析.

本书可作为各专业本科生、研究生数理统计或应用统计课程的基础教材或实验教材, 也可作为从事数据统计分析研究人员、工程技术人员的工具书或参考读物. 本书整个教材的教学安排可考虑以1:3的比例安排上机时间. 具体教学内容可根据需要进行取舍, 具体可参考下表的安排课时:

教学内容	选取章节	课时安排
R 语言入门	第一章, 第二章, 附录B	12
探索性数据分析	第三章, 第四章	12
数据统计分析	第五章, 第六章, 第八章, 第九章	24
选讲内容	第七章	8
	第十章	8
	第十一章	8

本书编写过程中, 参考了大量的资料文献. 得到了华东师范大学金融与统计学院全体老师, 特别是终生教授茆诗松老师的支持. 我的学生巍晓玲参与了本书第四和第五章初稿的编写工作, 徐安察参与了本书第六和第七章初稿的编写工作, 于巧丽参与了本书第八和第九章初稿的编写工作, 岳映婕参与了本书第十一章初稿的编写工作, 上海师范大学的朱杰老师参与了本书第十章的编写工作和全书的校对工作. 在全书的编写过程中, 得到了高等教育出版社领导和研究生教育与学术著作分社王丽萍女士的关心和帮助, 在此一并提示感谢.

由于编者水平有限, 书中一定存在不足甚至错误之处, 欢迎读者不吝指正.

作者

2008 年 5 月

目录

内容介绍	II
前言	I
第一章 R介绍	1
§1.1 S语言与R	1
§1.2 R的特点	2
§1.3 R的资源	3
§1.4 R的安装与运行	3
1.4.1 R软件的安装、启动与关闭	3
1.4.2 R程序包的安装与使用	4
第一章习题	6
第二章 R的基本原理与核心	8
§2.1 R的基本原理	8
§2.2 R的在线帮助	10
§2.3 一个简短的R会话	13
§2.4 R的数据结构	19
2.4.1 R的对象与属性	19
2.4.2 浏览对象的信息	22

2.4.3 向量的建立	24
2.4.4 数组与矩阵的建立	34
2.4.5 数据框(data frame)的建立	42
2.4.6 列表(list)的建立	48
2.4.7 时间序列(ts)的建立	49
§2.5 数据的存储与读取	51
2.5.1 数据的存储	51
2.5.2 数据的读取	52
§2.6 R 的图形功能	57
2.6.1 绘图函数	58
2.6.2 低级绘图命令	60
2.6.3 绘图参数	62
2.6.4 一个实例	64
§2.7 R 编程	72
2.7.1 循环和向量化	73
2.7.2 用R写程序	74
2.7.3 编写你自己的函数	75
2.7.4 养成良好的编程习惯	78
第二章习题	79
 第三章 概率与分布	 81
§3.1 随机抽样	81
§3.2 排列组合与概率的计算	82
§3.3 概率分布	83
3.3.1 离散分布的分布律	83
3.3.2 连续分布的密度函数	85

§3.4 R中内嵌的分布	91
§3.5 应用: 中心极限定理	93
3.5.1 中心极限定理	93
3.5.2 渐近正态性的图形检验	93
3.5.3 举例	95
第三章习题	99
第四章 探索性数据分析	101
§4.1 常用分布的概率函数图	101
§4.2 直方图与密度函数的估计	110
4.2.1 直方图	110
4.2.2 核密度估计	110
§4.3 单组数据的描述性统计分析	112
4.3.1 单组数据的图形描述	112
4.3.2 单组数据的描述性统计	117
§4.4 多组数据的描述性统计分析	120
4.4.1 两组数据的图形概括	120
4.4.2 多组数据的图形描述	126
4.4.3 多组数据的描述性统计	129
4.4.4 分组数据的图形概括	133
§4.5 分类数据的描述性统计分析	140
4.5.1 列联表的制作	140
4.5.2 列联表的图形描述	144
第四章习题	147
第五章 参数估计	150
§5.1 矩法估计和极大似然估计	150

5.1.1 矩法估计	150
5.1.2 极大似然估计	153
§5.2 单正态总体参数的区间估计	156
5.2.1 均值 μ 的区间估计	156
5.2.2 方差 σ^2 的区间估计	161
§5.3 两正态总体参数的区间估计	162
5.3.1 均值差 $\mu_1 - \mu_2$ 的置信区间	162
5.3.2 两方差比 σ_1^2/σ_2^2 的置信区间	166
§5.4 单总体比率 p 的区间估计	168
§5.5 两总体比率差 $p_1 - p_2$ 的区间估计	171
§5.6 样本容量的确定	173
5.6.1 估计正态总体均值时样本容量的确定	173
5.6.2 估计比例 p 时样本容量的确定	176
第四章习题	177
第六章 参数的假设检验	179
§6.1 假设检验与检验的 p 值	179
6.1.1 假设检验的概念与步骤	179
6.1.2 检验的 p 值	182
§6.2 单正态总体参数的检验	182
6.2.1 均值 μ 的假设检验	182
6.2.2 方差 σ^2 的检验: χ^2 检验	186
§6.3 两正态总体参数的检验	187
6.3.1 均值的比较: t 检验	187
6.3.2 方差的比较: F 检验	189
§6.4 成对数据的 t 检验	190

§6.5 单样本比率的检验	193
6.5.1 比率 p 的精确检验	193
6.5.2 比率 p 的近似检验	194
§6.6 两样本比率的检验	196
第六章习题	199
第七章 非参数的假设检验	200
§7.1 单总体位置参数的检验	200
7.1.1 中位数的符号检验	201
7.1.2 Wilcoxon符号秩检验	203
§7.2 分布的一致性检验: χ^2 检验	205
§7.3 两总体的比较与检验	209
7.3.1 χ^2 独立性检验	209
7.3.2 Fisher精确检验	211
7.3.3 Wilcoxon秩和检验法和Mann-Whitney U检验	213
7.3.4 Mood检验	215
§7.4 多总体的比较与检验	218
7.4.1 位置参数的Kruskal-Wallis秩和检验	218
7.4.2 尺度参数的Ansari-Bradley检验	220
7.4.3 尺度参数的Fligner-Killeen检验	221
第七章习题	223
第八章 方差分析	226
§8.1 单因子方差分析	226
8.1.1 数学模型	226
8.1.2 均值的多重比较	229
8.1.3 同时置信区间: Tukey法	233

8.1.4 方差齐性检验	236
§8.2 双因子方差分析	239
8.2.1 无交互作用的方差分析	239
8.2.2 有交互作用的方差分析	243
§8.3 协方差分析	248
第八章习题	254
第九章 回归分析与相关分析	260
§9.1 相关性及其度量	260
9.1.1 相关性概念	260
9.1.2 相关分析	261
§9.2 一元线性回归分析	264
9.2.1 数学模型	264
9.2.2 估计与检验	266
9.2.3 预测与控制	270
9.2.4 计算例子	272
§9.3 多元线性回归分析	275
9.3.1 数学模型	276
9.3.2 估计与检验	276
9.3.3 预测与控制	280
9.3.4 计算例子	281
§9.4 回归诊断	285
9.4.1 残差分析	286
9.4.2 影响分析	294
9.4.3 共线性诊断	298
§9.5 Logistic回归	301

第九章习题	309
第十章 多元统计分析介绍	315
§10.1 主成分分析与因子分析	315
10.1.1 主成分的简要定义与计算	316
10.1.2 主成分R通用程序	317
10.1.3 因子分析的简要定义与计算	320
10.1.4 因子分析R通用程序	322
§10.2 判别分析	325
10.2.1 距离判别	325
10.2.2 Fisher判别法	327
10.2.3 R通用程序	328
§10.3 聚类分析	332
10.3.1 基本思想	333
10.3.2 R通用程序	334
§10.4 典型相关分析	338
10.4.1 基本思想	338
10.4.2 R通用程序	340
§10.5 对应分析	343
10.5.1 基本思想	343
10.5.2 R通用程序	345
第十章习题	348
第十一章 贝叶斯统计分析	364
§11.1 贝叶斯统计分析与经典统计分析的比较	364
11.1.1 经典统计分析中存在的问题	364
11.1.2 对贝叶斯统计分析的质疑及褒奖	365

§11.2 贝叶斯统计分析与先验分布的选取	366
11.2.1 贝叶斯公式	367
11.2.2 先验分布的选取	369
11.2.3 贝叶斯分析体现了科学探索过程	371
§11.3 单参数贝叶斯统计分析	372
11.3.1 两项分布下的贝叶斯推断	372
11.3.2 正态分布下的贝叶斯统计推断	383
§11.4 多参数贝叶斯统计分析	388
11.4.1 方法概述	388
11.4.2 正态分布参数中的贝叶斯分析	388
11.4.3 随机模拟方法	389
11.4.4 一个实例	390
§11.5 分层贝叶斯统计分析	395
11.5.1 分层模型的建立及其贝叶斯推断	397
11.5.2 N-N模型与应用	400
§11.6 贝叶斯线性回归分析	408
11.6.1 模型的表示	408
11.6.2 后验分布	410
11.6.3 回归拟合	411
11.6.4 后验预测	411
第十一章习题	416
 附录 A 秩与结的介绍	 418
 附录 B R的图形界面	 420
§B.1 R Commander	420
B.1.1 功能	420

B.1.2 (网络)安装	420
B.1.3 运行	421
B.1.4 结构与使用	421
§B.2 PMG	422
B.2.1 功能	422
B.2.2 安装	423
B.2.3 结构与使用	424
附录 C R的编程环境	426
§C.1 R WinEdt	426
C.1.1 (网络)安装	426
C.1.2 运行	426
C.1.3 R WinEdt的特点	427
C.1.4 R WinEdt的菜单与热键	428
§C.2 Tinn-R	428
§C.3 SciViews R	429
参考文献	431

第一章 R介绍

本章概要

- ◇ R的功能与特点
- ◇ R的安装与运行
- ◇ R程序包的安装与运行

§1.1 S语言与R

R是一个有着强大统计分析及作图功能的软件系统，在GNU协议General Public Licence下免费发行，最先是由Ross Ihaka和Robert Gentleman共同创立，现在由**R**开发核心小组(R Development Core Team)维护，他们完全自愿、努力工作负责，并将全球优秀的统计应用软件打包提供给我们共享。

R可以看作是贝尔实验室(Bell Laboratories)的Rick Becker, John Chambers和Allan Wilks开发的**S**语言的一种实现或形式。因此，**R**是一种软件也可以说是一种语言。**S**语言现在主要内含在由Insightful公司经营的**S-PLUS**软件中。我们可以将**R**和**S-PLUS** 视为**S**语言的两种形式，**S/S-PLUS**方面的文档都可以直接用于**R**，不过**R**和**S**在设计理念上存在着许多不同，关于这方面的详细内容大家可以参考Ihaka & Gentleman (1996) 或随**R**同时发布的**R-FAQ**^[16]。本书今后主要使用**R**，有时也使用**R**软件、**R**语言或**R**系统来称呼这种形式的**S**语言。

§1.2 R的特点

现在越来越多的人开始接触、学习和使用**R**, 因为它有其显著的优点, 主要包括:

- 1) **免费**: 尽管**S-PLUS**是非常优秀的统计分析软件, 但你需要支付一笔费用, 而**R**是一个免费的统计分析软件(环境);
- 2) **浮点运算功能强大**: **R**可以作为一个高级科学计算器, 因为**R**同**Matlab**一样不需要编译就可执行代码;
- 3) **不依赖于操作系统**: **R**可以在运行于UNIX, Linux, Windows 和Macintosh 的操作系统上, 它们的安装文件以及安装说明都可以在CRAN (Comprehensive R Archive Network) 社区上下载;
- 4) **帮助功能完善**: **R**嵌入了一个非常实用的帮助系统—随软件所附的pdf或html帮助文件可以随时通过主菜单打开浏览或打印. 通过help命令可随时了解**R**所提供的各类函数的使用方法和例子;
- 5) **作图功能强大**: 其内嵌的作图函数能将产生的图片展示在一个独立的窗口中, 并能将之保存为各种形式的文件(例如jpg, png, bmp, ps, pdf, emf, pictex, xfig);
- 6) **统计分析能力尤为突出**: **R**内嵌了许多实用的统计分析函数, 统计分析的结果也能被直接显示出来, 一些中间结果(如p-值、回归系数、残差等)既可保存到专门的文件中, 也可以直接用于进一步的分析.

R的部分统计功能整合在**R**语言的底层, 但是大多数功能则以包的形式提供. 大约有25个包和**R**同时发布(被称为“标准”和“推荐”包), 更多的包可以通过网上或其CRAN 社区(<http://CRAN.R-project.org>) 得到, 它们都配有完整的pdf帮助文件, 且其版本会随**R**新版本的发行得到更新, 通过在线(或下载后)安装并加载后就可融入原来的**R**中, 实现有针对性的分析;

- 7) **可移植性强**:
 - **R**程序容易地移植到**S-PLUS**程序中; 反之**S-PLUS**的许多过程直接或稍作修改可用于**R**;
 - **R**与**Matlab**有许多相似的地方, 如都可作为高级计算器, 都可不经过编译直接运行源代码, 但是**R**侧重于统计分析, 而**Matlab**侧重于工程, 例如信号处理. 现在通过**R.Matlab**程序包可实现两者之间许多功能的共享, 具体见程序的说明.

- 许多常用的统计分析软件(如SPSS, SAS, Stata及EExcel)的数据文件都可读入R, 这样其它软件的数据或分析的中间结果可用于R, 并作出进一步的分析.
- 8) **较强大的拓展与开发能力:** R是开发新的交互式数据分析方法一个非常好的工具. 例如附录A介绍的R Commander就是一个非常成功的例子. 我们可以编制自己的函数来扩展现有的R语言, 或制作相对独立的统计分析包.
- 9) **灵活而不死板:** 一般的软件往往会直接展示分析的结果, 而R则将这些结果都存放在一个对象(object)里, 所以常常在分析执行结束后并不显示任何结果. 使用者(特别是初学者或非专业人员)可能会对此感到困惑, 其实这样的特点是非常有用的, 因为我们可以有选择地显示我们感兴趣的结果. 而有的软件(如SAS和SPSS)会同时显示几个窗口, 内容太多会使使用者无从选择和解释.

§1.3 R的资源

R的核心开发与维护小组通过R的主页, 即R工程(R Project)网站(<http://www.r-project.org>)及时发布有关信息, 包括R的简介、R的更新及宏包信息、R常用手册、已经出版的关于R的图书、R通讯和会议信息等. 你还可通过该主页预订邮件, 通过电子邮件发出求助或提供帮助.

R的CRAN社区是我们获得软件(及源代码)和资源的主要场所, 通过它或其镜像站点我们可以下载最新版本及大量的统计程序包(packages).

本书将使用Windows(95及以后)操作系统上的R, 其它操作系统上R的使用方法请参考R相关说明. 除R自带的运行平台R-GUI(R Graphic User's Interface)外, 本书附录A还提供了Windows下几个R的运行平台, R-Commander, R-WinEdt, R-Sciview和R-Tinn.

§1.4 R的安装与运行

1.4.1 R软件的安装、启动与关闭

R的安装: 从CRAN社区下载最新的封装好的R安装程序到本地计算机, 运行可执行的安装文件, 通常缺省的安装目录为C:\Program Files\R\R-x.x.x,

其中x.x.x为版本号. 安装时可以改变目录, 从2.2.0以后还可以选择中文作为基本语言, 这样RGui窗口的菜单都是中文的.

R的启动: 安装完成后点击桌面上R x.x.x快捷图标就可启动R的交互式用户窗口(R-GUI). R是按照问答的方式运行的, 也即你在命令提示符“>”后键入命令并回车, R就完成一些操作. 例如输入命令

```
> plot(rnorm(1000))
```

就可得到图1.1, 此命令的具体含义我们将在后面第二章叙述.

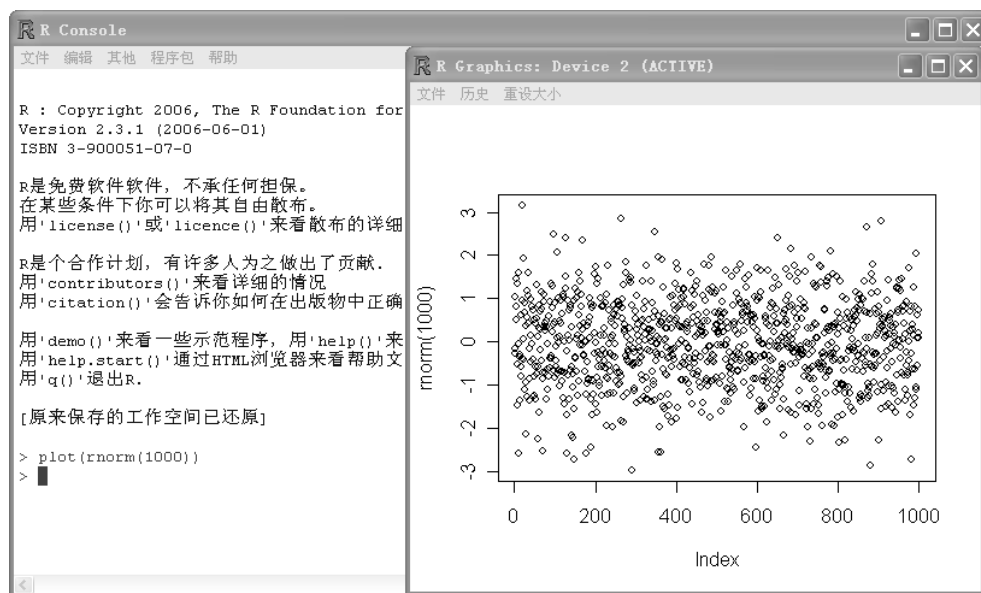


图 1.1 R的启动

R的退出: 在命令行键入q()或点击R-GUI右上角的叉叉. 退出时可选择保存工作空间, 缺省文件名为R安装目录的bin子目录下的R.RData. 以后可以通过命令load()或通过菜单“文件”下的“载入工作空间”加载, 进而继续你前一次的工作.

1.4.2 R程序包的安装与使用

R程序包的安装有三种方式:

- 1) **菜单方式:** 在已经联网的条件下, 按步骤“程序包⇒安装程序包...⇒选择CRAN镜像服务器⇒选定程序包”进行实时安装;

- 2) **命令方式**: 在已经联网的条件下, 在命令提示符后键入

```
> install.packages("PKname")
```

完成程序包PKname的安装.

- 3) **本地安装**: 在无上网条件下, 先从CRAN社区下载需要的程序包及与之关联的程序包, 再按第一种方式通过“程序包”菜单中的“用本机的zip文件安装程序包”选定本机上的程序包(zip文件)进行安装.

除R的标准程序包(如base包)外, 新安装的程序包在使用前必须先载入, 有两种载入方式:

- 1) **菜单方式**: 按步骤“程序包⇒载入程序包...”, 再从已有的程序包中选定需要的一个加载;
- 2) **命令方式**: 在命令提示符后键入

```
> library("PKname")
```

来加载程序包PKname.

若有必要, 我们还可通过步骤“程序包⇒更新程序包...”对本机的程序包进行实时更新.

注意: R命令对大小写敏感, 这在使用命令方式安装和载入程序包时应特别注意.

第一章习题

1.1 R与你学过的统计软件, 如SPSS, SAS, Matlab有何区别, 其主要的特点有哪些?

1.2 到CRAN社区(<http://cran.r-project.org/>)下载并安装R的最新(中文)版本, 并尝试R的启动与退化.

1.3 R可以作为一台很方便的计算器. 任取二个非零实数, 试用R完成它们的加、减、乘、除、乘方、开方、指数、对数等运算.

1.4 John Fox基于R开发了一套进行基础统计分析的菜单驱动的分析系统, 称为R Commander. 附录A介绍了一种菜单式的安装方法. 另一种是采用命令方式进行安装与加载, 其步骤为:

1) 用命令

```
> install.packages("Rcmdr")
```

来安装程序包Rcmdr(需要等待几分钟);

2) 再用命令

```
> load("Rcmdr")
```

加载程序包Rcmdr.

R Commander的结构与使用方法参见附录A的说明.

1.5 animation是由谢益辉建立的概率统计动态演示程序包, 请用命令或菜单的方法安装并加载animation, 并尝试下面的二个例子:

- 蒲丰投针试验:

```
> buffon.needle(nmax = 500, interval = 0)
```

- 中心极限定理:

```
> f = function(n) rchisq(n, 5)
> clt.ani(FUN = f)
```

具体使用方法参见程序包中的pdf说明文件.

1.6 登录R的社区主页<http://cran.r-project.org/>, 并进入左侧Software下的Packages, 浏览并感受R所提供的资源(程序包). 选择其中感兴趣的进行安装与试用, 例如概率统计教学演示程序包TeachingDemos和其在R Commander下的插件RcmdrPlugin.TeachingDemos.

第二章 R的基本原理与核心

本章概要

- ◇ R的基本原理
- ◇ R的求助方法
- ◇ R的主要数据结构
- ◇ R的图形功能
- ◇ R的编程方法

§2.1 R的基本原理

如第一章所述, 如果R已经安装在你的计算机中, 它就能立即运行一些可执行的命令了. R默认的命令提示符是‘>’, 它表示正在等待输入命令. 如果一个语句在一行中输不完, 按回车键, 系统会自动产生一个续行符“+”, 语句或命令输完后系统又会回到命令提示符. 在同一行中输入多个命令语句, 则需要使用分号来隔开. 在Windows系统中, 能直接运行下拉菜单中的一些操作命令(如在线帮助, 打开文件等, 见图1.1). 在学习一些R的命令之前, 让我们先了解R的基本工作原理.

首先, 同Matlab一样, R是一种编程语言, 但我们没有必要对此感到害怕, 因为R是一种解释性语言, 而不是编译语言, 也就意味着输入的命令能够直接被执行, 而不需要像其它语言(如C和FORTRAN)需要编译和连接等操作.

其次, R的语法非常简单和直观. 例如, 线性回归的命令`lm(y~x)`表示以 x 为自变量, y 为响应变量来拟合一个线性模型. 合法的R函数总是带有圆括号的形式, 即使括号内没有内容(如`ls()`). 如果直接输入函数名而不输入圆

括号, **R**则会自动显示该函数的一些具体内容. 因此在**R**中所有的函数后都带有圆括号以区别于对象(object). 当**R**运行时, 所有变量、数据、函数及结果都以对象的形式存入计算机的活动内存中, 并冠有相应的名字代号. 我们可以通过一些运算(如算术、逻辑、比较等)和一些函数(其本身也是对象)来对这些对象进行操作.

运行一个**R**函数可能不需要设定任何参量, 原因是所有的参量都可以被默认为缺省值, 当然也有可能该函数本身就不含任何参量.

再次, 在**R**中进行的所有操作都是针对存储在活动内存中的对象的. 数据、结果或图表的输入与输出都是通过对计算机硬盘中的文件读写而实现. 用户通过输入一些命令调用函数, 分析得出的结果可以被直接显示在屏幕上, 也可以存入某个对象或被写入硬盘(如图片对象). 因为产生的结果本身就是一种对象, 所以它们也能被视为数据并能像一般数据那样被处理分析. 数据文件即可从本地磁盘读取也可通过网络传输从远程服务器端获得.

最后, 所有能使用的**R**函数都被包含在一个库(library) 中, 该库存放在**R**安装文件夹的library目录下. 这个目录下含有具有各种功能的包(packages), 各个包也是按照目录的方式组织起来的. 其中名为base的包是**R**的核心, 因为它内嵌了**R**语言中所有像数据读写与操作这些最基本的函数. 在上述目录中的每个包内, 都有一个子目录**R**, 这个目录里又都含有一个与此包同名的文件, 该文件正是存放所有函数的地方.

R语言中最简单的命令莫过于通过输入一个对象的名字来显示其内容了. 例如, 一个名为n的对象, 其内容是数值10:

```
> n  
[1] 10
```

方括号中的数字1表示从n的第一个元素开始显示. 其实该命令的功能在这里与函数print()相似, 输出结果与print(n) 相同. 对象的名字必须是以一个字母开头(A-Z 或a-z), 中间可以包含字母、数字(0-9)、点(.)及下划线(_). 因为**R**对对象的名字区分大小写, 所以x和X就可以代表两个完全不同的对象.

一个对象可以通过赋值操作来产生, **R**语言中的赋值符号一般是由一个尖括号与一个负号组成的箭头形标志, 该符号可以是左到右的方向, 也可以相反. 赋值也可以用函数assign()实现, 还可以用等号“=”, 但它们很少使用. 例如

```
> n <- 10
> n
[1] 10
> 10 -> n
> n
[1] 10
> assign("n", 10)
> n
[1] 10
> n=10
> n
[1] 10
```

当然你也可以只是输入函数或表达式而不把它的结果赋给某个对象(如果这样在窗口中展示的结果将不会被保存到内存中), 这时我们就可将**R**作为一个计算器使用. 下面的例子说明了**R**中的算术运算符(加、减、乘、除、乘方、开方、指数)的使用方法.

```
> ((10 + 2) * 5-2^4)/4
[1] 13

> sqrt(3)+exp(-2)
[1] 1.867386
```

更为常用的是常量、向量、矩阵、数组等其它对象的赋值与运算, 我们将在后面讲述.

所有的高级语言都有注释语句, **R**中使用井号(#)表示注释的开始.

§2.2 R的在线帮助

学习一门编程语言离不开语句、函数和编程的语法和语义, **R**中的程序包都是由大量的进行统计分析的函数, 它们的含义和使用方法对于熟练使用**R**进行数据分析是至关重要的. 在此我们将**R**的帮助分成两类:

- 1) 关于**R**的基本知识: 通过命令


```
> help.start( )
```

或R用户界面上的“帮助”菜单的“html帮助”得到.

- i. **R**的常见问题(FAQ): 系统提供了二个版本, 其一为“R FAQ”, 其二为“R for Windows FAQ”, 它们随**R**的新版本同时发布与更新, 内容包括**R**的特点、安装、使用、界面、编程规则等.
- ii. **R**帮助手册, 也随新版本发布与更新, 共有6本手册: An Introduction to R, R Reference Manual, R Data input/output, R Language Definition, Writing R Extensions, R Installation and Administration. “帮助”菜单提供了它们的PDF电子版本, 便于打印. 初学者可看一下其中的第一本.

2) 关于**R**中的函数或关键字符:

i. 命令

```
> help(fun)
```

或

```
> ?fun
```

会立即显示名为“**fun**”函数的帮助页面, 而命令

```
> help("char")
```

则会显示某个具有特殊语法意义字符“**char**”的帮助页面. 页面的第一行一般会显示此函数或字符的所属的程序包(**package**), 然后是标题, 标题下面则是一些详细信息:

Description: brief description.

Usage: for a function, gives the name with all its arguments and the possible options (with the corresponding default values); for an operator gives the typical use.

Arguments: for a function, details each of its arguments.

Details: detailed description.

Value: if applicable, the type of object returned by the function or the operator.

See Also: other help pages close or similar to the present one.

Examples: some examples which can generally be executed without opening the help with the function **example**.

默认状态下, 函数`help()`只会在被载入内存的程序包中搜索. 选项`try.all.packages` 在缺省值是`FALSE`, 但如果把它设为`TRUE`, 则可在所有已安装的程序包中进行搜索. 如果读者确实想打开这样的页面而所属程序包又没有被载入内存时, 可以使用`package`这个选项. 请读者试试下面的两个命令.

```
> help("bs", try.all.packages=TRUE)
> help("bs", package = "splines")
```

ii. 命令

```
> apropos(fun)
```

或

```
> apropos("fun")
```

找出所有在名字中含有指定字符串“fun”的函数, 但只会在被载入内存中的程序包中进行搜索.

注意: 如果“fun”不是完整的函数名, 则前者会出错;

iii. 命令

```
> help.search("char")
```

列出所有在帮助页面含有字符“char”的函数, 它的搜索范围比`apropos("fun")`更广;

iv. 命令

```
> find(fun)
```

或

```
> find("fun")
```

得到名为“fun”函数所在的程序包;

v. 命令

```
> args(fun)
```

或

```
> args("fun")
```

得到名为“fun”函数的自变量列表.

对初学者而言, 帮助中例子(Examples)部分的信息是很有用的. 而仔细阅读自变量(Arguments)中的一些说明也是非常有必要的. 帮助中还包含了其它一些说明部分, 如注释(Notes), 参考文献(References)或作者(Author(s))等.

§2.3 一个简短的R会话

下面通过一个具体的例子来说明如何利用R软件进行数据的统计分析, 此例使用R内嵌的数据集mtcars. 它在datasets(数据)包中, 此包像base一样随R的启动自动加载.

数据的描述

命令

```
> ?mtcars
```

显示为

— ?mtcars的结果 —

```
mtcars           package:datasets           R Documentation

Motor Trend Car Road Tests

Description:
  The data was extracted from the 1974 _Motor Trend_
  US magazine, and comprises fuel consumption and 10
  aspects of automobile design and performance for 32
  automobiles (1973-74 models).

Usage:
  mtcars

Format:
  A data frame with 32 observations on 11 variables.

  [, 1]  mpg   Miles/(US) gallon
  [, 2]  cyl   Number of cylinders
  [, 3]  disp  Displacement (cu.in.)
  [, 4]  hp    Gross horsepower
  [, 5]  drat  Rear axle ratio
```

```
[, 6]  wt      Weight (lb/1000)
[, 7]  qsec    1/4 mile time
[, 8]  vs      V/S
[, 9]  am      Transmission (0 = automatic, 1 = manual)
[,10]  gear    Number of forward gears
[,11]  carb    Number of carburetors
```

它告诉我们数据集`mtcars`的基本信息. 它是美国Motor Trend收集的1973到1974年期间总共32辆汽车的11个指标: 油耗及10个与设计及性能方面的指标.

数据的浏览与编辑

1) 数据的浏览

- 命令

```
> mtcars
```

可以显示数据集`mtcars`中全部的32个观测值.

- 命令

```
> head(mtcars)
```

仅显示数据集`mtcars`中前7个观测值.

- 命令

```
> names(mtcars)
```

仅显示数据集`mtcars`中的变量, 在此为11个指标.

2) 数据的编辑

数据的编辑主要有两种方式(函数):

- 命令

```
> data.entry(mtcars)
```

通过R的数据编辑器打开数据集`mtcars`, 除了浏览数据集外, 这里我们还可以对变量及其观测值进行修改.

- 命令

```
> MTcars <- edit(mtcars)
```

同样启动R的数据编辑器, 在此可对原来的数据集`mtcars`进行编辑, 完成后将生成的新的数据集赋给`MTcars`, 而原来的数据集保持不变. 如果你要修改原来的数据集, 使命令`edit()`前后的数据集同名即可. 因此命令`edit(mtcars)`将无法完成对数据的修改. 命令

```
> xnew <- edit(data.frame( ))
```

可以编辑生成新的数据集`xnew`。另外, 对于一维的数据, `edit()`打开的是R Editor。试比较下面的例子中两个命令的区别

```
> x <- c(10.4, 5.6, 3.1, 6.4, 21.7)
```

```
> x
```

```
[1] 10.4  5.6  3.1  6.4 21.7
```

```
> data.entry(x)
```

```
> edit(x)
```

- 命令

```
> fix(mtcars)
```

可以完成数据集`mtcars`的直接修改。因此它等价于命令

```
> mtcars <- edit(mtcars)
```

注意:

- 1) 使用上面的三个命令将挂起R的对话窗口(R Console), 关闭编辑器即可继续进行R的对话。
- 2) 我们这里说的数据集就是下面一小节要讲的数据框(data frame)。数据对象中除了上面已经出现的向量和数据框外, 下面一节还要讲矩阵、数组、和列表。命令`data.entry()`和`edit()`都可用于编辑向量、矩阵、数据框和列表, 前者启用的都是R的数据编辑器, 后者有所不同: 对于向量、列表和数组`edit()`启用的是R Editor。
- 3) 尽管我们在R中可以浏览与编辑数据集`mtcars`, 但它们还无法对此数据集进行操作(分析), 例如命令

```
> mpg
```

无法看到变量`mpg`(每加仑公里数)的具体数值。这时我们需要激活或挂接(`attach`)数据集`mtcars`。命令

```
> attach(mtcars)
```

就激活`mtcars`, 使之成为当前的数据集。这时通过命令

```
> mpg
```

就可浏览变量`mpg`的32个值, 其它分析我们将在后面进行。

属性数据的分析

变量cyl(汽缸数)为属性变量, 命令

```
> table(cyl)
```

告诉我们变量cyl取3个值: 4, 6, 8, 相应的频数为11, 7, 14. 而命令

```
> barplot(table(cyl))
```

显示了cyl的频数直方图. 要注意的是, 命令

```
> barplot(cyl)
```

在此不适用, 它仅适用于数值型变量.

数值型数据的分析

统计分析中主要涉及数值型数据. 对此我们可考查它们的图形特征及常用的特征量.

- 画茎叶图(stem-and-leaf plot), 命令为

```
> stem(mpg).
```
- 画直方图, 命令为

```
> hist(mpg).
```
- 画框须图(stem-and-leaf plot), 命令为

```
> boxplot(mpg).
```
- 计算平均值, 命令为

```
> mean(mpg).
```
- 计算截去10%的平均值, 命令为

```
> mean(mpg, trim = .1).
```
- 按分组变量cyl计算mpg的分组平均值, 命令为

```
> tapply(mpg, cyl, mean)
```
- 计算cyl为4的那些mpg的平均值, 命令为

```
> mean(mpg[cyl == 4]).
```

- 计算四分位数的极差(`interquartile range`), 命令为
`> IQR(mpg)`.
- 计算样本常用的分位数: 极小、极大、中位数及两个四分位数, 命令为
`> quantile(mpg)`
或者
`> fivenum(mpg)`
- 计算由向量`prob`给定的各概率处的样本分位数, 命令为

```
> quantile(mpg, probs)
```

例如`probs = c(0.1, 0.5, 99.5)/100`. 可见, `quantile()`比`fivenum()`更为一般.

- 计算常用的描述性统计量, 它们分别是最小值(Min.)、第一四分位数(1st Qu.)、中位数(Median)、平均值(Mean)、第三分位数(3rd Qu.)和最大值(Max.), 命令为
`> summary(mpg)`.
- 计算标准差, 命令为
`> sd(mpg)`.
- 计算中位绝对离差(median absolute deviation), 命令为
`> mad(mpg)`.

寻找二元关系

- 画二维散点图, 例如`cyl`与`mpg`的散点图, 可通过下面的命令得到.

```
> plot(cyl,mpg)
```

注意: 相仿命令

```
plot(hp,mpg)
```

可得到`hp`与`mpg`的散点图. 但32个点对应了不同的汽缸, 因此按`cyl`为图例作出散点图更清晰, 命令为

```
> plot(hp,mpg,pch=cyl)
```

```
> legend(250,30,pch=c(4,6,8),
```

```
> legend=c("4 cylinders","6 cylinders","8 cylinders"))
```

- 拟合线性回归, 例如命令

```
> z <- lm(cyl ~ mpg)
```

可以得到

Call:

```
lm(formula = cyl ~ mpg)
```

Coefficients:

(Intercept)	mpg
11.2607	-0.2525

线性回归的截距为11.2607, 斜率为-0.2525.

- 相关系数(或 R^2)考查回归拟合好坏的程度. 命令

```
> cor(cyl,mpg)
```

可以得到相关系数(Pearson correlation coefficient) R , 其平方

```
> cor(cyl,mpg)^2
```

得到 R^2 为0.72618, 表明数据变化的72.6%可以用汽缸数(cyl)与每加仑的英里数(mpg)来刻画.

- 残差分析:

```
> lm.res <- lm(cyl ~ mpg)      # 将回归分析的结果作为对象
                                # 保存到lm.res中
> lm.resids <- resid(lm.res)   # 提取残差向量
> plot(lm.resids)              # 考查残差的散点图
> hist(lm.resids)              # 考查残差的直方图: 钟型?
> qqnorm(lm.resids)           # 残差的QQ图是否落在直线上?
```

结论: 从残差分析我们可以得出汽车的汽缸数与每加仑的里程数可以用线性回归来刻画.

结束分析并退出R

```
> detach(mtcars)              # 从内存中清除数据集mtcars
> q( )                        # 退出R
```


§2.4 R的数据结构

2.4.1 R的对象与属性

我们已经知道R通过一些对象来运行，这些对象是用它们的名称和内容来刻画的，其次也通过对象的数据类型即属性来刻画。所有的对象都有两个内在属性：类型和长度。类型是对象元素的基本种类，共有四种：

- 数值型, 包括
 - 整型
 - 单精度实型
 - 双精度实型
- 字符型
- 复数型¹
- 逻辑型(FALSE、TRUE或NA)

虽然还存在其它的类型，例如函数或表达式，但是它们并不能用来表示数据；长度是对象中元素的数目。对象的类型和长度可以分别通过函数`mode()`和`length()`得到。例如

```
> x <- 1
> mode(x)
[1] "numeric"
> length(x)
[1] 1
> A <- "Gomphotherium"; compar <- TRUE; z <- 1i
> mode(A); mode(compar); mode(z)
[1] "character"
[1] "logical"
[1] "complex"
```

无论什么类型的数据，缺失数据总是用NA(Not Available的意思)来表示；对很大的数值则可用指数形式表示：

¹本书不讨论复数型

```
> N <- 2.1e23
> N
[1] 2.1e+23
```

R可以正确地表示无穷的数值, 如用`Inf`和`-Inf`表示 $+\infty$ 和 ∞ , 或者用`NaN`(Not a Number 的意思)表示不是数字的值.

```
> x <- 5/0
> x
[1] Inf
> exp(x)
[1] Inf
> exp(-x)
[1] 0
> Inf - Inf
[1] NaN
> 0/0
[1] NaN
> sqrt{-7)
[1] NaN
Warning message:
产生了NaNs in: sqrt(-17)
> sqrt(-17+0i) # 按照复数进行运算
[1] 0+4.123106i
```

字符型的值输入时须加上双引号", 如果需要引用双引号的话, 可以让它跟在反斜杠“\”后面, 在某些函数如`cat()`的输出显示或`write.table()`写入磁盘时会被以特殊的方式处理. 例如

```
> x <- "Double quotes \" delimitate R's strings."
> x
[1] "Double quotes \" delimitate R's strings."
> cat(x)
Double quotes " delimitate R's strings.
```

另一种表示字符型变量的方法, 即用单引号(')来界定变量, 这种情况下不需要用反斜杠来引用双引号.

```
> x <- 'Double quotes " delimitate R\'s strings.'
```

```
> x
```

```
[1] "Double quotes \" delimitate R's strings."
```

表2.1概括了表示数据对象的类别:

表 2.1 数据对象及类型

对象	类型	是否允许 同一个对象中 有多种类型?
向量	数值型, 字符型, 复数型, 逻辑型	否
因子	数值型, 字符型	否
数组	数值型, 字符型, 复数型, 逻辑型	否
矩阵	数值型, 字符型, 复数型, 逻辑型	否
数据框	数值型, 字符型, 复数型, 逻辑型	是
时间序列(ts)	数值型, 字符型, 复数型, 逻辑型	否
列表	数值型, 字符型, 复数型, 逻辑型, 函数, 表达式, ...	是

说明

- 1) 向量是一个变量(的取值), 是**R**中最常用、最基本的操作对象; 因子是一个分类变量; 数组是一个 k 维的数据表; 矩阵是数组的一个特例, 其维数 $k = 2$.

注意: 数组或者矩阵中的所有元素都必须是同一种类型的; 数据框是由一个或几个向量和(或)因子构成, 它们必须是等长的, 但可以是不同的数据类型; “ts”表示时间序列数据, 它包含一些额外的属性, 例如频率和时间; 列表可以包含任何类型的对象, 包括列表!

- 2) 对于一个向量, 用它的类型和长度足够描述数据; 而其它的对象则另需一些额外信息, 这些信息由外在的属性给出, 例如这些属性中的表示对象

维数的dim. 比如一个2行2列的的矩阵, 它的dim是一对数值[2,2], 但是其长度是4.

- 3) **R**中有三种主要类型的运算符, 表2.2是这些运算符的列表. 其中数学运算符和比较运算符作用于两个元素上(例如 $x + y$, $a < b$); 数学运算符不只是作用于数值型或复数型变量, 也可以作用在逻辑型变量上; 在后一种情况中, 逻辑型变量被强制转换为数值型. 比较运算符可以适用于任何类型: 结果是返回一个或几个逻辑型变量; 逻辑型运算符适用于一个(对于“!”运算符)或两个逻辑型对象(对于其它运算符), 并且返回一个(或几个)逻辑性变量. 运算符“逻辑与”和“逻辑或”存在两种形式: “&”和“|”作用在对象中的每一个元素上并且返回和比较次数相等长度的逻辑值; “&&”和“||”只作用在对象的第一个元素上.

表 2.2 运算符

数学运算		比较运算		逻辑运算	
+	加法	<	小于	! x	逻辑非
-	减法	>	大于	x & y	逻辑与
*	乘法	<=	小于或等于	x && y	同上
/	除法	>=	大于或等于	x y	逻辑或
^	乘方	==	等于	x y	同上
%%	模	!=	不等于	xor(x, y)	异或
%%/	整除				

2.4.2 浏览对象的信息

函数ls()的功能是显示所有在内存中的对象. ls()只会列出对象名, 例如:

```
> name <- "Carmen"; n1 <- 10; n2 <- 100; m <- 0.5
> ls()
[1] "m"      "n1"     "n2"     "name"
```

如果只要显示出在名称中带有某个指定字符的对象，则通过设定选项`pattern`来实现(可简写为`pat`)):

```
> ls(pat = "m")
[1] "m"      "name"
```

如果进一步限定显示名称中以某个字母开头的对象，则可使用命令:

```
> ls(pat = "^m")
[1] "m"
```

运行函数`ls.str()`将会显示内存中所有对象的详细信息:

```
> ls.str()
m :   num 0.5 n1 :   num 10 n2 :   num 100 name :  chr "Carmen"
```

在`ls.str()`函数中另一个非常有用的选项是`max.level`, 它将规定显示有关对象信息的详细级别. 缺省情况下, `ls.str()` 将会列出关于对象的所有信息, 包括数据框、矩阵, 或数据列表的详细信息, 显示结果可能会很长. 但如果设定`max.level = -1` 就可以避免这种情况了. 试比较:

```
> M <- data.frame(n1, n2, m)
> ls.str(pat = "M")
M : `data.frame`:      1 obs. of  3 variables:
  $ n1: num 10
  $ n2: num 100
  $ m : num 0.5
> ls.str(pat="M", max.level=-1)
M : `data.frame`:      1 obs. of  3 variables:
```

要在内存中删除某个对象, 可利用函数`rm()`. 例如

- 运行`rm(x)`将会删除对象`x`
- 运行`rm(x,y)`将会删除对象`x`和`y`
- 运行`rm(list=ls())`则会删除内存中的所有对象
- 运行`rm(list=ls(pat="^m"))`则会删除对象中以字母`m`开头的对象

下面我们通过具体的例子说明向量(包括数值型向量、字符型向量、逻辑型向量和因子型向量)、矩阵、数据框、列表和时间序列的构成方法.

2.4.3 向量的建立

数值型向量的建立

统计分析中最为常用的是数值型的向量, 它们可用下面的四种函数建立:

- 1) `seq()` 或 “:” # 若向量(序列)具有较为简单的规律
- 2) `rep()` # 若向量(序列)具有较为复杂的规律
- 3) `c()` # 若向量(序列)没有什么规律
- 4) `scan()` # 通过键盘逐个输入

例子

```
> 1:10
[1] 1 2 3 4 5 6 7 8 9 10
> 1:10-1
[1] 0 1 2 3 4 5 6 7 8 9
> 1:(10-1)
[1] 1 2 3 4 5 6 7 8 9       # 注意括号有无的区别
> z <- seq(1,5,by=0.5)      # 等价于 seq(from=1,to=5,by=0.5)
> z
[1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
> z <- seq(1,10,length=11)  # 等价于 seq(1,10,length.out=11)
> z
[1] 1.0 1.9 2.8 3.7 4.6 5.5 6.4 7.3 8.2 9.1 10.0
> z <- rep(2:5,2)           # 等价于 rep(2:5, times=2)
> z
[1] 2 3 4 5 2 3 4 5
> z <- rep(2:5,rep(2,4))
[1] 2 2 3 3 4 4 5 5
> z <- rep(1:3, times = 4, each = 2)
> z
```

```

[1] 1 1 2 2 3 3 1 1 2 2 3 3 1 1 2 2 3 3 1 1 2 2 3 3
> z <- x<-c(42,7,64,9)
> z
[1] 42  7 64  9
> z <- scan( )           # 通过键盘建立向量
1: 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
10:
Read 9 items
> z
[1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
> z <- sequence(3:5)
> z
[1] 1 2 3 1 2 3 4 1 2 3 4 5
> z <- sequence(c(10,5))
> z
[1]  1  2  3  4  5  6  7  8  9 10  1  2  3  4  5

```

字符型向量的建立

字符和字符向量在R中广泛使用,比如图表的标签.在显示的时候,相应的字符串由双引号界定,字符串在输入时可以使用单引号(')或双引号(").引号(")在输入时应当写作\".字符向量可以通过函数c()连接.函数paste()可以接受任意个参数,并从它们中逐个取出字符并连成字符串,形成的字符串的个数与参数中最长字符串的长度相同.如果参数中包含数字的话,数字将被强制转化为字符串.在默认情况下,参数中的各字符串是被一个空格分隔的,不过通过参数sep=string 用户可以把它更改为其他字符串,包括空字符串.例如

```

> Z <- c("green","blue sky","-99")
> Z
[1] "green"    "blue sky" "-99"
> labs <- paste(c("X","Y"), 1:10, sep="")
> labs
[1] "X1"  "Y2"  "X3"  "Y4"  "X5"  "Y6"  "X7"  "Y8"  "X9"  "Y10"

```

逻辑型向量的建立

与数值型向量相同，**R**允许对逻辑向量进行操作. 一个逻辑向量的值可以是TRUE, FALSE和NA. 前两个通常简写为T和F². 逻辑向量是由条件给出的. 譬如

```
> x <- c(10.4, 5.6, 3.1, 6.4, 21.7)
> temp <- x > 13
> temp
[1] FALSE FALSE FALSE FALSE TRUE
```

temp为一个与x长度相同, 元素根据是否与条件相符而由TRUE或FALSE组成的向量. 逻辑向量可以在普通的运算中被使用, 此时它们将被转化为数字向量, FALSE当做0, 而TRUE当做1. 再看几个简单的例子:

```
> 7!=6
[1] TRUE
> !(7==6)
[1] TRUE
> !(7==6)==1
[1] TRUE
> (7==9)|(7>0)
[1] TRUE
> (7==9)&(7>0)
[1] FALSE
```

因子型向量的建立

一个因子或因子向量不仅包括分类变量本身, 还包括变量不同的可能水平(即使它们在数据中不出现). 因子利用函数factor()创建. factor()的调用格式如下:

factor() 的调用格式

```
factor(x, levels = sort(unique(x), na.last = TRUE),
      labels = levels, exclude = NA, ordered = is.ordered(x))
```

²注意T和F仅仅是默认被指向TRUE和FALSE的变量, 而不是系统的保留字.

说明: `levels` 用来指定因子的水平(缺省值是向量`x`中不同的值); `labels`用来指定水平的名字; `exclude`表示从向量`x`中剔除的水平值; `ordered`是一个逻辑型选项, 用来指定因子的水平是否有次序. 这里`x`可以是数值型或字符型, 这样对应的因子也就称为数值型因子或字符型因子. 因此, 因子的建立可以通过字符型向量或数值型向量来建立, 且可以转化.

1) 将字符型向量转换成因子

```
> a <- c("green", "blue", "green", "yellow")
> a <- factor(a)
a
[1] green blue green yellow
Levels: blue green yellow
```

2) 将数值型向量转换成因子

```
> b <- c(1,2,3,1)
> b <- factor(b)
> b
[1] 1 2 3 1
Levels: 1 2 3
```

3) 将字符型因子转换为数值型因子

```
> a <- c("green", "blue", "green", "yellow")
> a <- factor(a)
> levels(a)<-c(1,2,3,4)
> a
[1] 2 1 2 3
Levels: 1 2 3 4
> ff <- factor(c("A", "B", "C"), labels=c(1,2,3))
> ff
[1] 1 2 3
Levels: 1 2 3
```

4) 将数值型因子转换为字符型因子

```

> b <- c(1,2,3,1)
> b <- factor(b)
> levels(b) <- c("low", "middle", "high")
> b
[1] low    middle high    low
Levels: low middle high
> ff <- factor(1:3, labels=c("A", "B", "C"))
ff
[1] A B C
Levels: A B C

```

注: 函数`levels()`用来提取一个因子中可能的水平值, 例如

```

> ff <- factor(c(2, 4), levels=2:5)
> ff
[1] 2 4
Levels: 2 3 4 5
> levels(ff)
[1] "2" "3" "4" "5"

```

- 5) 函数`gl()`能产生规则的因子序列. 这个函数的用法是`gl(k,n)`, 其中`k`是水平数, `n`是每个水平重复的次数. 此函数有两个选项: `length`用来指定产生数据的个数, `label`用来指定每个水平因子的名字. 例如:

```

> gl(3, 5)
[1] 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3
Levels: 1 2 3
> gl(3, 5, length=30)
[1] 1 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 1 1 1 1 1 2 2 2 2 2 3 3 3 3 3
Levels: 1 2 3
> gl(2, 6, label=c("Male", "Female"))
[1] Male    Male    Male    Male    Male    Male
[7] Female  Female  Female  Female  Female  Female
Levels: Male Female
> gl(2, 10)
[1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2

```

```
Levels: 1 2
> gl(2, 1, length=20)
[1] 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
Levels: 1 2
> gl(2, 2, length=20)
[1] 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2
Levels: 1 2
```

数值型向量的运算

向量可以用于算术表达式中, 操作是按照向量中的元素一个一个进行的. 同一个表达式中的向量并不需要具有相同的长度, 如果它们的长度不同, 表达式的结果是一个与表达式中最长向量有相同长度的向量, 表达式中较短的向量会根据它的长度被重复使用若干次(不一定是整数次), 直到与长度最长的向量相匹配, 而常数将被不断重复 — 这一规则称为循环法则(recycling rule). 例如, 命令

```
> x <- c(10.4, 5.6, 3.1, 6.4, 21.7)
> y <- c(x, 0, x)
> v <- 2*x + y + 1
```

产生一个长度为11的新向量 v , 其中 $2 * x$ 被重复2.2次, y 被重复1次, 常数1被重复11次. 为了方便使用, 我们对向量的运算稍作细分:

- 向量与一个常数的加、减、乘、除为向量的每一个元素与此常数进行加、减、乘、除;
- 向量的乘方($^$)与开方(`sqrt`)为每一个元素的乘方与开方, 这对象log, exp, sin, cos, tan 等普通的运算函数同样适用;
- 同样长度向量的加、减、乘、除等运算为对应元素进行加、减、乘、除等;
- 不同长度向量的加、减、乘、除遵从循环法则(recycling rule), 但要注意这种场合通常要求向量的长度为倍数关系, 否则会出现警告: “长向量并非短向量的整数倍”.

下面举例说明

```
> 5+c(4,7,17)
[1] 9 12 22
> 5*c(4,7,17)
[1] 20 35 85
> c(-1,3,-17)+c(4,7,17)
[1] 3 10 0
> c(2,4,5)^2
[1] 4 16 25
> sqrt(c(2,4,25))
[1] 1.414214 2.000000 5.000000
> 1:2+1:4
[1] 2 4 4 6
> 1:4+1:7
[1] 2 4 6 8 6 8 10
Warning message:
长的目标对象长度不是短的目标对象长度的整倍数 in: 1:4 + 1:7
```

常用统计函数

最后列出统计分析中常用的函数与作用(见表2.1).

图 2.1 统计分析中常用的函数与作用

统计函数	作用
<code>max(<i>x</i>)</code>	返回向量 <i>x</i> 中最大的元素
<code>min(<i>x</i>)</code>	返回向量 <i>x</i> 中最小的元素
<code>which.max(<i>x</i>)</code>	返回向量 <i>x</i> 中最大元素的下标
<code>which.min(<i>x</i>)</code>	返回向量 <i>x</i> 中最小元素的下标
<code>mean(<i>x</i>)</code>	计算样本(向量) <i>x</i> 的均值
<code>median(<i>x</i>)</code>	计算样本(向量) <i>x</i> 的中位数
<code>mad(<i>x</i>)</code>	计算中位绝对离差

<code>var(x)</code>	计算样本(向量) x 的方差
<code>sd(x)</code>	计算向量 x 的标准差
<code>range(x)</code>	返回长度为2的向量: <code>c(min(x), max(x))</code>
<code>IQR(x)</code>	计算样本的四分位数极差
<code>quantile(x)</code>	计算样本常用的分位数 ³
<code>summary(x)</code>	计算常用的描述性统计量(最小、最大、平均值、中位数和四分位数)
<code>length(x)</code>	返回向量 x 的长度
<code>sum(x)</code>	给出向量 x 的总和
<code>prod(x)</code>	给出向量 x 的乘积
<code>rev(x)</code>	取向量 x 的逆序
<code>sort(x)</code>	将向量 x 按升序排序, 选项 <code>decreasing=TRUE</code> 表示降序
<code>order(x)</code>	返回 x 的秩(升序), 选项 <code>decreasing=TRUE</code> 得到降序的秩
<code>rank(x)</code>	返回 x 的秩
<code>cumsum(x)</code>	返回向量 x 和累积和(其第 i 个元素是从 $x[1]$ 到 $x[i]$ 的和)
<code>cumprod(x)</code>	返回向量 x 和累积积(其第 i 个元素是从 $x[1]$ 到 $x[i]$ 的积)
<code>cummin(x)</code>	返回向量 x 和累积最小值(其第 i 个元素是从 $x[1]$ 到 $x[i]$ 的最小值)
<code>cummax(x)</code>	返回向量 x 和累积最大值(其第 i 个元素是从 $x[1]$ 到 $x[i]$ 的最大值)

³`quantile(x)`仅计算 x 的极小、极大、中位数及两个四分位数, 更一般地使用`quantile(x, probs)`可计算给定向量`probs`处的样本分位数.

<code>var(x, y)</code>	计算样本(向量) x 与 y 的协方差
<code>cov(x, y)</code>	计算样本(向量) x 与 y 的协方差
<code>cor(x, y)</code>	计算样本(向量) x 与 y 的相关系数
<code>outer(x, y)</code>	计算样本(向量) x 与 y 的外积 ⁴

函数`max()`, `min()`, `median()`, `var()`, `sd()`, `sum()`, `cumsum()`, `cumprod()`, `cummax()`, `cummin()`对于矩阵及数据框的意义有方向性. 对于矩阵, `cov()`和`cor()`分别用于求矩阵的协方差阵和相关系数阵, 这些将在后面举例说明.

向量的下标(index)与子集(元素)的提取

选择一个向量的子集(元素)可以通过在其名称后追加一个方括号中的索引向量来完成. 更一般地, 任何结果为一个向量的表达式都可以通过追加索引向量来选择其中的子集. 这样的索引向量有四种不同的类型.

- 1) 正整数向量 — 提取向量中对应的元素. 这种情况下索引向量中的值必须在集合 $\{1, 2, \dots, \text{length}(x)\}$ 中. 返回的向量与索引向量由相同的长度, 且按索引向量的顺序排列. 例如`x[6]`是 x 的第六个元素, 而
`> x[1:10]`选取了 x 的前10个元素(假设 x 的长度不小于10).
`> x[c(1,4)]`
 取出向量 x 的第1和第4个元素.
- 2) 负整数向量 — 去掉向量中与索引向量对应的元素. 例如
`> y <- x[-(1:5)]`
 从 x 中去除前5个元素得到 y .
- 3) 字符串的向量. 这种可能性只存在于拥有`names`属性并由它来区分向量中元素的向量. 这种情况下一个由名称组成的子向量起到了和正整数的索引向量相同的效果. 例如

⁴函数`outer()`的一般形式为`(x, y, "op")`, 其中`op`可为任一四则运算符.

```
> fruit <- c(5, 10, 1, 20)
> names(fruit) <- c("orange", "banana", "apple", "peach")
fruit
orange banana apple peach
      5      10      1      20
> lunch <- fruit[c("apple", "orange")]
> lunch
apple orange
      1      5
```

- 4) 逻辑的向量 — 取出满足条件的元素. 在索引向量中返回值是TRUE的元素所对应的元素将被选出, 返回值为FALSE的值所对应的元素将被忽略. 例如

```
> x <- c(42, 7, 64, 9)
> x > 10 # 值大于10的元素逻辑值
[1] TRUE FALSE TRUE FALSE
> x[x > 10] # 值大于10的元素
[1] 42 64
> x[x < 40 & x > 10]
numeric(0)
> x[x > 10] <- 10
> x
[1] 10 7 10 9
> y = runif(100, min=0, max=1) # (0,1)上100个均匀分布随机数
> sum(y < 0.5) # 值小于0.5的元素的个数
[1] 47
> sum(y[y < 0.5]) # 值小于0.5的元素的值的和
[1] 10.84767
> y <- x[!is.na(x)] # x中的非缺失值
> z <- x[(!is.na(x)) & (x > 0)] # x中的非负非缺失值
```

2.4.4 数组与矩阵的建立

前面已经指出数组是一个 $k(\geq 1)$ 维的数据表; 矩阵是数组的一个特例, 其维数 $k = 2$, 而上面所述的向量自然也可看成维数为 $k = 1$ 的数组⁵. 而且向量、数组或者矩阵中的所有元素都必须是同一种类型的. 对于一个向量, 其属性由其类型和长度构成; 而对于数组与矩阵, 除了类型和长度两个属性外, 还需要维数`dim`这个属性来描述. 因此如果一个向量需要在R中以数组的方式被处理, 则必须含有一个维数向量作为它的`dim`属性.

数组的建立

R中数组由函数`array()`建立, 其一般格式为:

```
> array(data, dim, dimnames)
```

其中`data`为一向量, 其元素用于构建数组; `dim`为数组的维数向量(为数值型向量); `dimnames`为由各维的名称构成的向量(为字符型向量), 缺省为空.

以一个3维的数据为例来说明. 设 A 是一个存放在向量 a 中的24个数据项组成的数组, A 的维数向量为`c(3,4,2)`. 维数可由命令

```
> dim(A) <- c(3,4,2)
```

建立. 这样, 命令

```
> A <- array(a, dim = c(3,4,2))
```

就建立了数组 A . 24个数据项在数组 A 中的顺序依次为: $a[1,1,1]$, $a[2,1,1]$, \dots , $a[2,4,2]$, $a[3,4,2]$. 我们再来看一个具体的例子:

```
> A <- array(1:8, dim = c(2, 2, 2))
```

```
> A
```

```
, , 1
```

```
  [,1] [,2]
```

```
[1,]    1    3
```

```
[2,]    2    4
```

```
, , 2
```

```
  [,1] [,2]
```

```
[1,]    5    7
```

```
[2,]    6    8
```

⁵通常使用`c()`建立向量, 使用`matrix()`建立矩阵, 使用`array()`建立数组, 因此它们在R中的属性是不同的


```
> dim(A)
[1] 2 2 2
> dimnames(A) <- list(c("a", "b"), c("c", "d"), c("e", "f"))
> A
, , e
  c d
a 1 3
b 2 4

, , f
  c d
a 5 7
b 6 8
> colnames(A)
[1] "c" "d"
> rownames(A)
[1] "a" "b"
> dimnames(A)
[[1]]
[1] "a" "b"
[[2]]
[1] "c" "d"
[[3]]
[1] "e" "f"
```

如果数据项太少，则采用循环准则填充数组(或矩阵)，见下面的第二个例子。

矩阵的建立

因为矩阵是数组的特例，因此矩阵也可以用函数`array()`来建立，例如

```
> A <- array(1:6, c(2,3))
> A
      [,1] [,2] [,3]
[1,]     1     3     5
```

```

[2,]    2    4    6
> A<-array(1:4,c(2,3))
> A
      [,1] [,2] [,3]
[1,]    1    3    1
[2,]    2    4    2
> A<-array(1:8,c(2,3))
> A
      [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6

```

然而, 由于矩阵在数学及统计中的特殊性, 在R中最为常用的是使用命令`matrix()`建立矩阵, 而对角矩阵用函数`diag()`建立更为方便, 例如

```

> X <- matrix(1, nr = 2, nc = 2)
      [,1] [,2]
[1,]    1    1
[2,]    1    1
> X <- diag(3)    # 生成单位阵
      [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
> v <- c(10, 20, 30)
> diag(v)
      [,1] [,2] [,3]
[1,]   10    0    0
[2,]    0   20    0
[3,]    0    0   30
> diag(2.5, nr = 3, nc = 5)
      [,1] [,2] [,3] [,4] [,5]
[1,]  2.5  0.0  0.0    0    0
[2,]  0.0  2.5  0.0    0    0
[3,]  0.0  0.0  2.5    0    0
> X <- matrix(1:4, 2)    # 等价于X <- matrix(1:4, 2, 2)

```

```
> X
      [,1] [,2]
[1,]    1    3
[2,]    2    4
> rownames(X) <- c("a", "b")
> colnames(X) <- c("c", "d")
> X
      c d
a 1 3
b 2 4
> dim(X)
[1] 2 2
> dimnames(X)
[[1]]
[1] "a" "b"
[[2]]
[1] "c" "d"
```

注意:

- 循环准则仍然适用于`matrix()`, 但要求数据项的个数等于矩阵的列数的倍数, 否则会出现警告.
- 矩阵的维数使用`c()`会得到不同的结果(除非是方阵), 因此需要小心.
- 数据项填充矩阵的方向可通过参数`byrow`来指定, 其缺省是按列填充的(`byrow=FALSE`). `byrow=TRUE`表示按行填充数据.

再看几个例子:

```
> X <- matrix(1:4, 2, 4) # 按列填充
> X
      [,1] [,2] [,3] [,4]
[1,]    1    3    1    3
[2,]    2    4    2    4
> X <- matrix(1:4, 2, 3)
Warning message:
```

```

In matrix(1:4, 2, 3) : 数据长度[4]不是矩阵列数[3]的整倍数
> X <- matrix(1:4, c(2, 3)) # 不经常使用
> X
      [,1] [,2]
[1,]     1     3
[2,]     2     4
> X <- matrix(1:4, 2, 4, byrow=TRUE) # 按行填充
> X
      [,1] [,2] [,3] [,4]
[1,]     1     2     3     4
[2,]     1     2     3     4

```

数组与矩阵的下标(index)与子集(元素)的提取

同向量的下标一样，矩阵与数组的下标可以使用正整数、负整数和逻辑表达式，从而实现子集的提取或修改。考查矩阵

```

x <- matrix(1:6, 2, 3)
> x
      [,1] [,2] [,3]
[1,]     1     3     5
[2,]     2     4     6

```

- 提取一个元素

```

> x[2,2]
[1] 4

```

- 提取若一个或若干个行或列⁶

```

> x[2,]
[1] 2 4 6
> x[,2]
[1] 3 4
> x[,2,drop=FALSE]

```

⁶R的缺省规则是返回一个维数尽可能低的对象，这可以通过修改选项drop的值来改变。

```

      [,1]
[1,]    3
[2,]    4
> x[,c(2,3),drop=FALSE]
      [,1] [,2]
[1,]    3    5
[2,]    4    6

```

- 去掉若一个或若干个行与列

```

> x[-1,]
[1] 2 4 6
> x[, -2]
      [,1] [,2]
[1,]    1    5
[2,]    2    6

```

- 添加与替换元素

```

> x[,3] <- NA
> x
      [,1] [,2] [,3]
[1,]    1    3   NA
[2,]    2    4   NA
> x[is.na(x)] <- 1 # 缺失值用1代替
> x
      [,1] [,2] [,3]
[1,]    1    3    1
[2,]    2    4    1

```

对矩阵的运算(函数)

对于矩阵的运算, 我们分通常的矩阵代数运算与统计运算来分别讨论.

1) 矩阵的代数运算:

- 转置函数`t()`:

```
> X <- matrix(1:6, 2, 3)
> X
      [,1] [,2] [,3]
[1,]    1    3    5
[2,]    2    4    6
> t(X)
      [,1] [,2]
[1,]    1    2
[2,]    3    4
[3,]    5    6
```

- 提取对角元diag():

```
> X <- matrix(1:4, 2, 2)
> diag(X)
[1] 1 4
```

- 几个矩阵按行合并rbind()与按列合并cbind():

```
> m1 <- matrix(1, nr = 2, nc = 2)
> m2 <- matrix(2, nr = 2, nc = 2)
> rbind(m1, m2)
      [,1] [,2]
[1,]    1    1
[2,]    1    1
[3,]    2    2
[4,]    2    2
> cbind(m1, m2)
      [,1] [,2] [,3] [,4]
[1,]    1    1    2    2
[2,]    1    1    2    2
```

- 矩阵的逐元乘积“*”:

```
> m2*m2
      [,1] [,2]
[1,]    4    4
[2,]    4    4
```

- 矩阵的代数乘积“%*%”:

```

> rbind(m1, m2) %*% cbind(m1, m2)
      [,1] [,2] [,3] [,4]
[1,]     2     2     4     4
[2,]     2     2     4     4
[3,]     4     4     8     8
[4,]     4     4     8     8
> cbind(m1, m2) %*% rbind(m1, m2)
      [,1] [,2]
[1,]    10    10
[2,]    10    10

```

- 方阵的行列式`det()`

```

> X<-matrix(1:4, 2)
> X
      [,1] [,2]
[1,]     1     3
[2,]     2     4
> det(X)
[1] -2

```

- 其它函数: 交叉乘积(cross product), 函数为`crossprod()`; 特征根与特征向量, 函数为`eigen()`; QR分解, 函数为`qr()`, 等等.

2) 矩阵的统计运算:

在讲述向量时我们已经提到过函数`max()`, `min()`, `median()`, `var()`, `sd()`, `sum()`, `cumsum()`, `cumprod()`, `cummax()`, `cummin()`对于矩阵(及数据框)有方向性. 而函数`cov()`和`cor()`分别用于计算矩阵的协方差阵和相关系数阵.

正是由于矩阵的排列是有方向性的, 在R中规定矩阵是按列排的, 若没有特别说明上述函数的使用也是按列计算的, 但也可以通过选项`MARGIN`来改变. 下面我们要用到对一个对象施加某种运算的函数`apply()`, 其格式为

```
> apply(X, MARGIN, FUN)
```

其中`X`为参与运算的矩阵, `FUN`为上面的一个函数或“+”、“-”、“*”、“\”(必须放在引号中), `MARGIN=1`表示按列计算, `MARGIN=2`表示按行计算, `MARGIN=c(1,2)`表示按行列计算(在至少3维的数组中使用).

我们还用到`sweep()`函数, 命令

```
> sweep(X, MARGIN, STATS, FUN)
```

表示从矩阵 X 中按 $MARGIN$ 计算 $STATS$, 并从 X 中除去(`sweep out`). 下面举几个例子加以说明:

- 求均值, 中位数等:

```
> m<-matrix(rnorm(n=12),nrow=3)
> apply(m, MARGIN=1, FUN=mean) # 求各行的均值
[1] -0.2540148  0.5474583  0.1493290
> apply(m, MARGIN=2, FUN=mean) # 求各列的均值
[1] -0.5389053  0.4731592  0.7821656 -0.1260561
```

- 标准化:

```
> scale(m, center=T, scale=T)
```

- 减去中位数:

```
> row.med <- apply(m, MARGIN=1, FUN=median)
> sweep(m, MARGIN=1, STATS=row.med, FUN="-")
```

2.4.5 数据框(data frame)的建立

统计分析中一个完整的数据集通常是由若干个变量的若干个观测值组成的, 在R中称为数据框. 数据框是一个对象, 它与前面讲的矩阵与二维数组形式上是类似的, 也是二维的, 也有维数这个属性, 且各个变量的观测值有相同的长度. 但不同的是: 在数据框中, 行与列的意义是不同的, 其中的列表示变量, 而行表示观测. 显示数据框时左侧会显示观测值的序号.

数据框的建立分为直接的与间接的两种方法:

数据框的直接建立

若你在R中建立了一些向量并试图想由它们生成数据框, 则可以使用函数`data.frame()`. 例如

```
> x=c(42,7,64,9)
> y=1:4
> z.df=data.frame(INDEX = y, VALUE = x)
```


	INDEX	VALUE
1	1	42
2	2	7
3	3	64
4	4	9

数据框中的向量必须有相同的长度或长度有倍数关系, 如果其中有一个比其它的短, 它将按循环法则“循环”整数次. 例如

```
> weight <- c(70.6, 56.4, 80, 59.5)
> x <- (c("adult", "teen", "adult", "teen"))
> wag <- data.frame(weight, age = x)
> wag
  weight  age
1  70.6 adult
2  56.4  teen
3  80.0 adult
4  59.5  teen
> x <- 1:4; y <- 2:4
> data.frame(x, y)
错误于data.frame(x, y) : 变元值意味着不同的行数 4, 3
```

数据框的间接建立

一个数据框还可以通过数据文件(文本文件、EXCEL文件或其它统计软件的数据文件)读取并建立, 在此我们仅通过一个例子来说明如何通过函数`read.table()`读取文件`c:\data\foo.txt`中的观测值, 并建立一个数据框. 其它间接方法可参考下一节“数据的存贮与读取”的介绍. 已知存于`foo.txt`上的数据如下:

	treat	weight
A	3.4	
B	NA	
A	5.8	

则下面的命令建立了数据框`foo`.

```
> foo <- read.table(file = "c:/data/foo.txt", header = T)
> foo
  treat weight
1     A    3.4
2     B     NA
3     A    5.8
```

适用于数据框的函数

在上一小节中我们所讨论的关于矩阵的统计计算函数`max()`、`min()`、`median()`、`var()`、`sd()`、`sum()`、`cumsum()`、`cumprod()`、`cummax()`、`cummin()`、`cov()`、`cor()`同样适用于数据框，意义也相同。这里通过R内嵌的另一个数据集Puromycin来说明`summary()`、`pairs()`和`xtable()`等的使用。

```
> attach(Puromycin) # 挂接数据集使之激活
> help(Puromycin)   # 显示前几行
> summary(Puromycin) # 显示主要的描述性统计量
```

conc	rate	state
Min. :0.0200	Min. : 47.0	treated :12
1st Qu.:0.0600	1st Qu.: 91.5	untreated:11
Median :0.1100	Median :124.0	
Mean :0.3122	Mean :126.8	
3rd Qu.:0.5600	3rd Qu.:158.5	
Max. :1.1000	Max. :207.0	

从summary可以看出，变量conc和rate是数值型的，而state为因子变量。变量之间的关系可以通过成对数据散点图考查：

```
> pairs(Puromycin, panel = panel.smooth)
```

最后使用xtabs()函数由交叉分类因子产生一个列联表：

```
> xtabs(~state + conc, data = Puromycin)
```

	conc
state	0.02 0.06 0.11 0.22 0.56 1.1

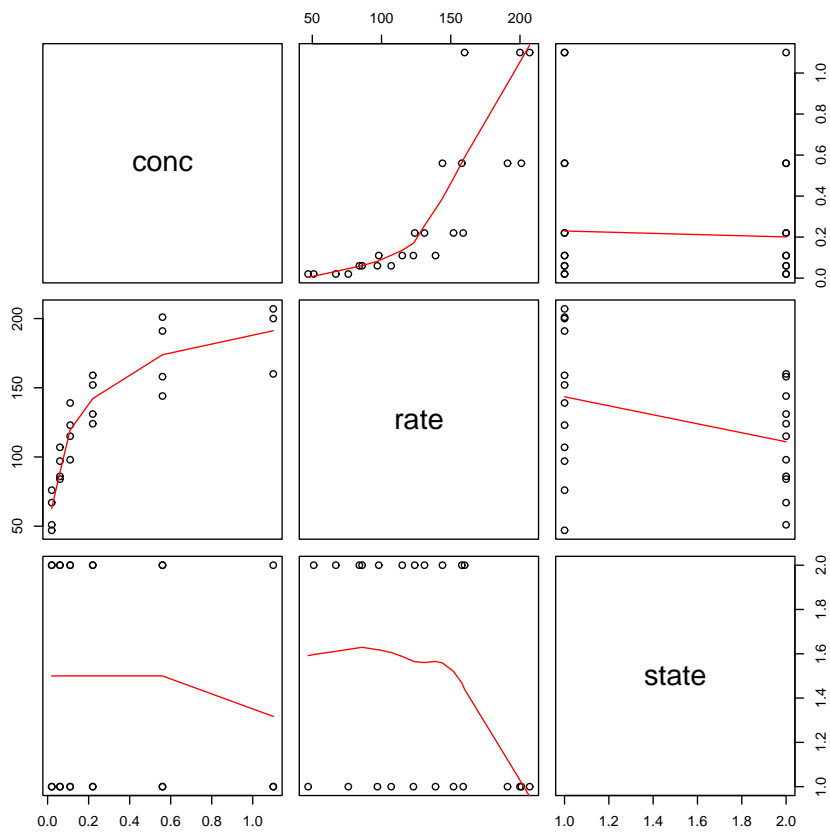


图 2.2 Puromycin的成对散点图

```
treated    2    2    2    2    2    2
untreated  2    2    2    2    2    1
```

数据框的下标与子集的提取

数据框的下标与子集的提取与矩阵基本相同. 不同的是: 对于列我们可以使用变量的名称, 仍以数据集**Puromycin**进行举例说明.

- 提取单个元素

```
> Puromycin[1, 1]
[1] 0.02
```

- 提取一个子集, 例如第1, 3, 5行, 第1, 3列

```
> Puromycin[c(1, 3, 5), c(1, 3)]
  conc  state
1 0.02 treated
3 0.06 treated
5 0.11 treated
> Puromycin[c(1, 3, 5), ]
  conc rate  state
1 0.02   76 treated
3 0.06   97 treated
5 0.11  123 treated
```

常使用变量名称来指定列的位置, 上面的命令等价于

```
> Puromycin[c(1, 3, 5), c("conc", "state")]]
```

- 提取一列(变量的值). 一个数据框的变量对应了数据框的一列, 如果变量有名称, 则可直接使用“数据框名\$变量名”这种格式指向对应的列. 例如

```
> Puromycin$conc      # 等价于 Puromycin[,1]
[1] 0.02 0.02 0.06 0.06 0.11 0.11 0.22 0.22 0.56 0.56
[11] 1.10 1.10 0.02 0.02 0.06 0.06 0.11 0.11 0.22 0.22
[21] 0.56 0.56 1.10
> Puromycin$state
[1] treated   treated   treated   treated   treated
[6] treated   treated   treated   treated   treated
[11] treated   treated   untreated untreated untreated
[16] untreated untreated untreated untreated untreated
[21] untreated untreated untreated
Levels: treated untreated
```

- 提取满足条件的子集

```
> subset(Puromycin, state == "treated" & rate > 160)
  conc rate  state
9 0.56  191 treated
```

```
10 0.56 201 treated
11 1.10 207 treated
12 1.10 200 treated
> subset(Puromycin, conc > mean(conc))
      conc rate    state
9  0.56  191   treated
10 0.56  201   treated
11 1.10  207   treated
12 1.10  200   treated
21 0.56  144 untreated
22 0.56  158 untreated
23 1.10  160 untreated
```

数据框中添加新变量

在原有的数据框中添加新的变量有三种方法. 假设我们想在Puromycin中增加变量iconc, 其定义为 $1/\text{conc}$, 则可分别使用:

1) 基本方法

```
> Puromycin$iconc <- 1/Puromycin$conc
```

2) 使用with() 函数

```
> Puromycin$iconc <- with(Puromycin, 1/conc)
```

3) 使用transform()函数, 且可一次性定义多个变量

```
> Puromycin <- transform(Puromycin, iconc = 1/conc,
  sqrtconc = sqrt(conc))
> head(Puromycin)
      conc rate    state    iconc  sqrtconc
1 0.02    76 treated 50.00000 0.1414214
2 0.02    47 treated 50.00000 0.1414214
3 0.06    97 treated 16.66667 0.2449490
4 0.06   107 treated 16.66667 0.2449490
5 0.11   123 treated  9.09091 0.3316625
6 0.11   139 treated  9.09091 0.3316625
```

2.4.6 列表(list)的建立

复杂的数据分析时, 仅有向量与数据框还不够, 有时需要生成包含不同类型的对象. **R**的列表(list)就是包含任何类型的对象.

列表可以用函数`list()`创建, 方法与创建数据框类似(见§2.4.5). 和`data.frame()`一样, 缺省值没有给出对象的名称. 列表的下标与子集的提取也与数据框没有本质区别. 数据分析时通常是在提取部分对象后按上面讲述的向量、矩阵或数据框等运算进行, 在此不再一一列举. 下面仅举一例进行说明.

```
> L1 <- list(1:6, matrix(1:4, nrow = 2))
> L1
[[1]]
[1] 1 2 3 4 5 6

[[2]]
      [,1] [,2]
[1,]     1     3
[2,]     2     4

L2 <- list(x = 1:6, y = matrix(1:4, nrow = 2))
> L2
$x
[1] 1 2 3 4 5 6

$y
      [,1] [,2]
[1,]     1     3
[2,]     2     4

> L2$x
[1] 1 2 3 4 5 6
> L2[1]
$x
[1] 1 2 3 4 5 6
```

```
> L2[[1]]
[1] 1 2 3 4 5 6

> L2[[1]][2]
[1] 2
> L2$x[2]
[1] 2
> L2$y[4]
[1] 4
```

2.4.7 时间序列(ts)的建立

由函数`ts()`通过一向量或者矩阵创建一个一元的或多元的时间序列(time series), 它称为`ts`型对象, 其调用格式为:

函数`ts()`的调用格式

```
ts(data = NA, start = 1, end = numeric(0), frequency = 1,
    deltat = 1, ts.eps = getOption("ts.eps"), class, names)
```

函数`ts()`可带一些表明序列特征的选项(其本身可使用缺省值), 它们是:

<code>data</code>	一个向量或者矩阵
<code>start</code>	第一个观察值的时间, 为一个数字或者是一个由两个整数构成的向量(参见下面的例子)
<code>end</code>	最后一个观察值的时间, 指定方法和 <code>start</code> 相同
<code>frequency</code>	单位时间内观察值的频数(频率)
<code>deltat</code>	两个观察值间的时间间隔(例如, 月度数据的取值为1/12); <code>frequency</code> 和 <code>deltat</code> 必须并且只能给定其中的一个
<code>ts.eps</code>	序列之间的误差限. 如果序列之间的频率差异小于 <code>ts.eps</code> , 则认为这些序列的频率相等.
<code>class</code>	对象的类型. 一元序列的缺省值是" <code>ts</code> ", 多元序列的缺省值是 <code>c("mts", "ts")</code>
<code>names</code>	一个字符型向量, 给出多元序列中每个一元序列的名称, 缺省为 <code>data</code> 中每列数据的名称或者 <code>Series 1, Series 2, ...</code>

我们看几个用`ts()`创建时间序列的一些例子:

```

> ts(1:10, start = 1959)
Time Series:
Start = 1959
End = 1968
Frequency = 1
[1] 1 2 3 4 5 6 7 8 9 10
> ts(1:47, frequency = 12, start = c(1959, 2))
      Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
1959      1  2  3  4  5  6  7  8  9 10 11
1960 12 13 14 15 16 17 18 19 20 21 22 23
1961 24 25 26 27 28 29 30 31 32 33 34 35
1962 36 37 38 39 40 41 42 43 44 45 46 47
> ts(1:10, frequency = 4, start = c(1959, 2))
      Qtr1 Qtr2 Qtr3 Qtr4
1959      1  2  3
1960  4  5  6  7
1961  8  9 10
> ts(matrix(rpois(36,5),12,3), start=c(1961,1),frequency=12)
      Series 1 Series 2 Series 3
Jan 1961      8      5      4
Feb 1961      6      6      9
Mar 1961      2      3      3
Apr 1961      8      5      4
May 1961      4      9      3
Jun 1961      4      6     13
Jul 1961      4      2      6
Aug 1961     11      6      4
Sep 1961      6      5      7
Oct 1961      6      5      7
Nov 1961      5      5      7
Dec 1961      8      5      2

```

本书不讨论时间序列的统计分析, 有兴趣的可参考Zivot与Wang(2002).

§2.5 数据的存储与读取

对于在文件读取和写入的工作，**R**使用工作目录来完成。如果一个文件不在工作目录里则必须给出它的路径。可以使用命令`getwd()`(获得工作目录)来找到目录，使用命令`setwd("C:/data")`将当前的工作目录改变为C:\data(注意**R**命令中目录的分割符使用正斜杠“/”或两个反斜杠“\\”)。工作目录的设置也可通过“文件”菜单的“改变当前目录...”来完成⁷。

2.5.1 数据的存储

保存为文本文件

R软件中使用函数`write.table()`或`save()`在文件中写入一个对象，一般是写一个数据框，也可以是其它类型的对象(向量、矩阵、数组、列表等)。我们以数据框为例加以说明，例如数据框`d`是用下面的命令建立的：

```
> d <- data.frame(obs = c(1, 2, 3), treat = c("A", "B", "A"),
  weight = c(2.3, NA, 9))
```

1) 保存为简单的文本文件

```
> write.table(d, file = "c:/data/foo.txt",
  row.names = F, quote = F)
```

其中选项`row.names = F`表示行名不写入文件，`quote = F`表示变量名不放在双引号中。

2) 保存为逗号分割的文本文件

```
> write.csv(d, file = "c:/data/foo.csv",
  row.names = F, quote = F)
```

3) 保存为**R**格式文件

```
> save(d, file = "c:/data/foo.Rdata")
```

在经过了一段时间的分析后，常需要将工作空间的映像保存起来，命令为

⁷如果不设定工作目录，在读写文件时也可将目录直接写在`file`参数中。

```
> save.image( )
```

实际上它等价于

```
> save(list =ls(all=TRUE), file=".RData")
```

我们了也可通过菜单“文件”下的“保存工作空间”来完成. 上述三个函数的选项及具体使用请查看它们的帮助文件.

2.5.2 数据的读取

文本文件数据的读取

R可以用下面的函数读取存储在文本文件(ASCII)中的数据: `read.table()`, `scan()`和`read.fwf()`.

1) 使用函数`read.table()`

函数`read.table()`用来创建一个数据框, 所以它是读取表格形式的数据的主要方法, 这一点我们在前一节已经提到. 我们再举一个例子, 先在“c:\data”下建立文件houses.dat, 其内容为

	Price	Floor	Area	Rooms	Age	Cent.heat
01	52.00	111.0	830	5	6.2	no
02	54.75	128.0	710	5	7.5	no
03	57.50	101.0	1000	5	4.2	no
04	57.50	131.0	690	6	8.8	no
05	59.75	93.0	900	5	1.9	yes

则使用命令:

```
> setwd("C:/data")
> HousePrice <- read.table(file="houses.dat")
```

建立数据框HousePrice. 默认情况下, 数值项(除了行标号)将被当作数值变量读入. 非数值变量, 如例子中的Cent.heat, 将被作为因子读入. 如果明确数据的第一行作为表头行, 则使用header选项:

```
> HousePrice <- read.table("houses.dat", header=TRUE)
```

除上面的基本形式外, `read.table()` 还有4个变形: `read.csv()`, `read.csv2()`, `read.delim()`, `read.delim2()`. 前二个读取用逗号分割的数据; 后二个则针对使用其它分割符分割的数据(它们不使用行号). 具体可参考`read.table()`的帮助文件. 如果上面的文件在取消行号后每一个数据项后加上逗号“,”, 并改名为`house.csv`, 则上述命令改为

```
> HousePrice <- read.csv("houses.csv", header=TRUE)
```

2) 使用函数`scan()`

函数`scan()`比`read.table()`要更加灵活, 它们的区别之一是: `scan()`可以指定变量的类型, 例如我们先建立文件`C:\data\data.dat`:

```
M      65    168
M      70    172
F      54    156
F      58    163
```

命令:

```
> mydata <- scan("data.dat", what = list("", 0, 0))
```

读取了文件`data.dat`中三个变量, 第一个是字符型变量, 后两个是数值型变量. 其中第二个参数是一个名义列表结构, 用来确定要读取的三个向量的模式. 在名义列表中, 我们可以直接命名对象. 例如

```
> mydata <- scan("data.dat",
+ what = list(Sex="", Weight=0, Height=0))
> mydata
$Sex
[1] "M" "M" "F" "F"

$Weight
[1] 65 70 54 58

$Height
[1] 168 172 156 163
```

另一个重要的区别在于`scan()`可以用来创建不同的对象: 向量、矩阵、数据框、列表等. 在缺省情况下(即`what`被省略), `scan()`将创建一个数值型向量. 如果读取的数据类型与缺省类型或指定类型不符, 则将返回一个错误信息. 更一般的说明可参考`scan()`的帮助文件.

3) 使用函数`read.fwf()`

函数`read.fwf()`可以用来读取文件中一些固定宽度格式的数据. 除了选项`widths`用来说明读取字段的宽度外, 其它选项与`read.table()`基本相同. 例如, 我们先建立文件`C:\data\data.txt`:

```
A1.501.2
A1.551.3
B1.601.4
B1.651.5
C1.701.6
C1.751.7
```

命令:

```
> mydata <- read.fwf("data.txt", widths=c(1, 4, 3),
                     col.names=c("X", "Y", "Z"))
```

得到

```
      X      Y      Z
1 A 1.50 1.2
2 A 1.55 1.3
3 B 1.60 1.4
4 B 1.65 1.5
5 C 1.70 1.6
6 C 1.75 1.7
```

更详细的说明可参考`read.fwf()`的帮助文件.

Excel数据的读取

有两种简单的方法获得Excel电子表格中的数据.

1) 利用剪贴板

一种简单不过的方法是打开Excel中电子表格, 选中需要的数据区域, 再复制到剪贴板中(使用CTRL+C). 然后在R中键入命令

```
> mydata <- read.delim("clipboard")
```

2) 使用程序包RODBC.

要得到文件"c:\data\body.xls"中工作表1(sheet1)中的数据, 设为

Sex	Weight	Height
M	65	168
M	70	172
F	54	156
F	58	163

可以使用命令

```
> library(RODBC)
> z <- odbcConnectExcel("c:/data/body.xls")
> foo <- sqlFetch(z, "Sheet1")
> close(z)
```

R中数据集的读取

1) R的标准数据datasets

R提供了—个基本的数据集包datasets, 其中包含了100多个数据集(通常为数据框和列表). 它随着R的启动全部—次性自动载入, 通过命令

```
> data( )
```

就可列出全部的数据集(包括已经通过library()加载的其它程序包的数据集). 输入数据集的名字或用help(dataname)就可看到你所关心的数据集的信息.

2) 专用程序包中的数据集

要读取其他已经安装的专用程序包中的数据, 可以使用package参数, 例如

```
> data(package="pkname") # pkname 为已安装的程序包的名字
```

就可以列出程序包pkname中的所有数据集, 但要注意的是它们还未被载入到R系统中供浏览. 而命令

```
> data(dataname, package="pkname")
```

则载入程序包pkname中的名为dataname的数据集. 这时数据dataname的信息就可通过其名字或help()进行浏览. 用户发布的程序包是一个丰富的数据集来源.

注意:

- 从上面的例子我们看到data()有两个功能: 浏览数据列表和加载数据集, 但可浏览到的数据集并不一定已经加载;
- 命令library()用于加载程序包, 程序包加载后其函数可以使用, 但其中的数据集仍未载入, 仍需要使用data()加载. 因此通常的做法是逐个使用下面的命令

```
> library("pkname")  
> data( ) # 或 data(package="pkname")  
> data(dataname) # data(dataname, package="pkname")
```

- data(dataname)将从第一个能够找到data(dataname)的程序包中载入这个数据集. 为避免载入同名的其它数据集, 加上package选项是有必要的.
- 加载的数据集中的变量是不能直接按其名字参与运算的, 例如在R刚启动后, 数据集mtcars中的变量mpg是无法直接按其名字浏览与参与计算的. 例如要计算其平均值, 可以使用命令

```
> mean(mtcars$mpg)
```

得到20.09062. 另一个方法是使用命令attach(mtcars)将此数据集挂接进来, 成为当前的数据集. 这时R就将这个数据集中的变量放到一个临时的目录中供访问. 这时与上面命令等价的是

```
> attach(mtcars)  
> mean(mpg)  
[1] 20.09062
```

一个好的习惯是在不用此数据集时将它挂起(卸载,detach):

```
> detach(mtcars)
```

R格式的数据

R的数据或更为一般的对象(包括向量、数据框、列表、函数等)可以通过`save()`保存起来, 文件名以**Rdata**为后缀. 例如我们将**mtcars**中的变量**mpg**和**hp**生成为数据框**mtcars2**, 并保存在文件**myR.Rdata**中:

```
> attach(mtcars)
> mtcars2 <- data.frame(mtcars[,c(1,4)])
> save(mtcars2, "c:/data/myR.Rdata")
```

而命令

```
> load("c:/data/myR.Rdata")
```

则可以重新加载进来. 涉及多个数据集的统计分析经常使用这种方法保存与加载数据.

其它统计软件数据的读取

R也可以读取其它统计软件的数据文件(如SAS, SPSS, Stata, S-PLUS)和访问SQL类型的数据库, 程序包**foreign**提供了这一便利. 由于它们仅对**R**的高级应用有用, 我们在此不再细说, 具体可参考随机**R**同时发行的**R data Import/Export**手册.

§2.6 R 的图形功能

R提供非常多样的绘图功能. 我们可以通过**R**提供的二组演示例子进行了解:

- `demo(graphics)`为二维的图形示例;
- `demo(persp)`为三维的图形示例.

我们在这里不可能详细说明R软件在绘图方面的所有功能，主要是因为每个绘图函数都有大量的选项，使得图形的绘制十分的灵活多变。

绘图函数的工作方式与本文前面描述的工作方式大为不同，不能把绘图函数的结果赋给一个对象⁸，其结果将直接输出到一个“绘图设备”上。绘图设备是一个绘图的窗口或是一个文件。

在R中有两种绘图函数：

- 1) 高级绘图函数(*high-level plotting functions*)创建一个新的图形
- 2) 低级绘图函数(*low-level plotting functions*)在现存的图形上添加元素。

另外绘图参数(*graphical parameters*)提供了丰富的绘图选项，可以使用缺省值或者用函数`par()`修改。更高级的图形可使用`grid`和`lattice`绘图包实现，具体可查看其中的说明文档。Paul Murrell(2006)系统地介绍了R中作图方法和例子。

2.6.1 绘图函数

表2.3概括了R中的高级绘图函数。

图 2.3 高级绘图函数

函数名	功能
<code>plot(x)</code>	以 x 的元素值为纵坐标、以序号为横坐标绘图
<code>plot(x, y)</code>	x (在 x -轴上)与 y (在 y -轴上)的二元作图
<code>sunflowerplot(x, y)</code>	同上，但是以相似坐标的点作为花朵，其花瓣数目为点的个数
<code>pie(x)</code>	饼图
<code>boxplot(x)</code>	盒形图(“box-and-whiskers”)
<code>stripchart(x)</code>	把 x 的值画在一条线段上，样本量较小时可作为盒形图的替代
<code>coplot(x~y z)</code>	关于 z 的每个数值(或数值区间)绘制 x 与 y 的二元图

⁸有一些值得注意的例外：`hist()`和`barplot()`仍然把生成的数据结果作为列表或矩阵。

<code>interaction.plot (f1, f2, y)</code>	如果 f1 和 f2 是因子，作 y 的均值图，以 f1 的不同值作为 x 轴，而 f2 的不同值对应不同曲线；可以用选项 fun 指定 y 的其他的统计量(缺省计算均值， fun=mean)
<code>matplot(x,y)</code>	二元图，其中 x 的第一列对应 y 的第一列， x 的第二列对应 y 的第二列，依次类推。
<code>dotchart(x)</code>	如果 x 是数据框，作Cleveland点图(逐行逐列累加图)
<code>fourfoldplot(x)</code>	用四个四分之一圆显示2times2列联表情况(x 必须是 dim=c(2, 2, k) 的数组，或者是 dim=c(2, 2) 的矩阵，如果 k = 1)
<code>assocplot(x)</code>	Cohen-Friendly图，显示在二维列联表中行、列变量偏离独立性的程度
<code>mosaicplot(x)</code>	列联表的对数线性回归残差的马赛克图
<code>pairs(x)</code>	如果 x 是矩阵或是数据框，作 x 的各列之间的二元图
<code>plot.ts(x)</code>	如果 x 是类" ts "的对象，作 x 的时间序列曲线， x 可以是多元的，但是序列必须有相同的频率和时间
<code>ts.plot(x)</code>	同上，但如果 x 是多元的，序列可有不同的时间但须有相同的频率
<code>hist(x)</code>	x 的频率直方图
<code>barplot(x)</code>	x 的值的条形图
<code>qqnorm(x)</code>	正态分位数一分位数图
<code>qqplot(x, y)</code>	y 对 x 的分位数一分位数图
<code>contour(x, y, z)</code>	等高线图(画曲线时用内插补充空白的值)， x 和 y 必须为向量， z 必须为矩阵，使得 dim(z)=c(length(x), length(y)) (x 和 y 可以省略)
<code>filled.contour (x, y, z)</code>	同上，等高线之间的区域是彩色的，并且绘制彩色对应的值的图例
<code>image(x, y, z)</code>	同上，但是实际数据大小用不同色彩表示
<code>persp(x, y, z)</code>	同上，但为透视图
<code>stars(x)</code>	如果 x 是矩阵或者数据框，用星形和线段画出
<code>symbols(x, y, ...)</code>	在由 x 和 y 给定坐标画符号(圆，正方形，长方形，星，温度计式或者盒形图)，符号的类型、大小、颜色等由另外的变量指定
<code>termplot(mod.obj)</code>	回归模型(mod.obj)的(偏)影响图

R的绘图函数的部分选项是一样的. 下面列出主要的共同选项及其缺省值:

选项	功能
<code>add=FALSE</code>	如果是TRUE, 叠加图形到前一个图上(如果有的话)
<code>axes=TRUE</code>	如果是FALSE, 不绘制轴与边框
<code>type="p"</code>	指定图形的类型, "p": 点, "l": 线, "b": 点连线, "o": 同上, 但是线在点上, "h": 垂直线, "s": 阶梯式, 垂直线顶端显示数据, "S": 同上, 但是在垂直线底端显示数据
<code>xlim=, ylim=</code>	指定轴的上下限, 例如 <code>xlim=c(1, 10)</code> 或者 <code>xlim=range(x)</code>
<code>xlab=, ylab=</code>	坐标轴的标签, 必须是字符型值
<code>main=</code>	主标题, 必须是字符型值
<code>sub=</code>	副标题(用小字体)

2.6.2 低级绘图命令

R的低级作图命令作用于现存的图形上的, 下表给出了一些主要的:

函数名	功能
<code>points(x, y)</code>	添加点(可以使用选项 <code>type=</code>)
<code>lines(x, y)</code>	同上, 但是添加线
<code>text(x, y, labels, ...)</code>	在 (x,y) 处添加用 <code>labels</code> 指定的文字; 典型的用法是: <code>plot(x, y, type="n"); text(x, y, names)</code>
<code>mtext(text, side=3, line=0, ...)</code>	在边空添加用 <code>text</code> 指定的文字, 用 <code>side</code> 指定添加到哪一边(参照下面的 <code>axis()</code>); <code>line</code> 指定添加的文字距离绘图区域的行数
<code>segments(x0, y0, x1, y1)</code>	从 (x_0,y_0) 各点到 (x_1,y_1) 各点画线段

<code>arrows(x0, y0, x1, y1, angle= 30, code=2)</code>	同上, 但加画箭头. 如果 <code>code=2</code> , 则在各 <code>(x0,y0)</code> 处画箭头; 如果 <code>code=1</code> , 则在各 <code>(x1,y1)</code> 处画箭头; 如果 <code>code=3</code> , 则在两端都画箭头 <code>angle</code> 控制箭头轴到箭头边的角度.
<code>abline(a,b)</code>	绘制斜率为 <code>b</code> 和截距为 <code>a</code> 的直线
<code>abline(h=y)</code>	在纵坐标 <code>y</code> 处画水平线
<code>abline(v=x)</code>	在横坐标 <code>x</code> 处画垂直线
<code>abline(lm.obj)</code>	画由 <code>lm.obj</code> 确定的回归线
<code>rect(x1, y1, x2, y2)</code>	绘制长方形, <code>(x1, y1)</code> 为左下角, <code>(x2,y2)</code> 为右上角
<code>polygon(x, y)</code>	绘制连接各 <code>x,y</code> 坐标确定的点的多边形
<code>legend(x, y, legend)</code>	在点 <code>(x,y)</code> 处添加图例, 说明内容由 <code>legend</code> 给定
<code>title()</code>	添加标题, 也可添加一个副标题
<code>axis(side, vect)</code>	画坐标轴. <code>side=1</code> 时画在下边; <code>side=2</code> 时画在左边; <code>side=3</code> 时画在上边; <code>side=4</code> 时画在右边. 可选参数 <code>at</code> 指定画刻度线的位置坐标
<code>box()</code>	在当前的图上加上边框
<code>rug(x)</code>	在 <code>x</code> -轴上用短线画出 <code>x</code> 数据的位置
<code>locator(n, type="n", ...)</code>	在用户用鼠标在图上点击 <code>n</code> 次后返回 <code>n</code> 次点击的坐标 <code>(x, y)</code> ; 并可以在点击处绘制符号(<code>type="p"</code> 时)或连线(<code>type="l"</code> 时), 缺省情况下不画符号或连线

注意: 用

```
> text(x, y, expression(...))
```

可以在一个图形上加上数学公式, 函数`expression`把自变量转换为数学公式. 例如,

```
> text(x, y, expression(p==over(1,1+e^-(beta*x+alpha))))
```

在图中相应坐标点 (x, y) 处显示下面的方程:

$$p = \frac{1}{1 + e^{-(\beta x + \alpha)}}.$$

为了能在表达式中代入某个变量的值,我们可以使用函数`substitute()`和`as.expression()`。例如,为了代入 R^2 的值(之前计算并储存在对象`Rsquared`中)

```
> text(x, y, as.expression(substitute(R^2==r, list(r=Rsquared))))
```

在图中相应坐标点 (x, y) 处显示:

$$R^2 = 0.9856298.$$

如果只显示3位小数,上述命令修改为:

```
> text(x, y, as.expression(substitute(R^2==r,
                                     list(r=round(Rsquared, 3)))))
```

它将显示:

$$R^2 = 0.986.$$

最后,用斜体字显示 R^2 ,命令为

```
> text(x, y, as.expression(substitute(italic(R)^2==r,
                                     list(r=round(Rsquared, 3)))))
```

得到

$$R^2 = 0.986.$$

2.6.3 绘图参数

除了低级作图命令之外,图形的显示也可以用绘图参数来改良.绘图参数可以作为图形函数的选项(但不是所有参数都可以这样用),也可以用函数`par()`来永久地改变绘图参数,也就是说后来的图形都将按照函数`par()`指定的参数来绘制.例如,下面的命令:

```
> par(bg="yellow")
```

将导致后来的图形都以黄色的背景来绘制.有73个绘图参数,其中一些有非常相似的功能.这些参数详细的列表可以通过`help(par)`获得.下面的表格只列举了最常用的参数.

参数	功能
adj	控制关于文字的对齐方式: 0是左对齐, 0.5是居中对齐, 1是右对齐, 值> 1时对齐位置在文本右边的地方, 取负值时对齐位置在文本左边的地方; 如果给出两个值(例如c(0, 0)), 第二个只控制关于文字基线的垂直调整
bg	指定背景色(例如bg="red", bg="blue"; 用colors()可以显示657种可用的颜色名)
bty	控制图形边框形状, 可用的值为: "o", "l", "7", "c", "u" 和"]" (边框和字符的外表相像); 如果bty="n"则不绘制边框
cex	控制缺省状态下符号和文字大小的值; 另外, cex.axis控制坐标轴刻度数字大小, cex.lab控制坐标轴标签文字大小, cex.main控制标题文字大小, cex.sub控制副标题文字大小
col	控制符号的颜色; 和cex类似, 还可用: col.axis, col.lab, col.main, col.sub
font	控制文字字体的整数(1: 正常, 2: 斜体, 3: 粗体, 4: 粗斜体); 和cex类似, 还可用: font.axis, font.lab, font.main, font.sub
las	控制坐标轴刻度数字标记方向的整数(0: 平行于轴, 1: 横排, 2: 垂直于轴, 3: 竖排)
lty	控制连线的线型, 可以是整数(1: 实线, 2: 虚线, 3: 点线, 4: 点虚线, 5: 长虚线, 6: 双虚线), 或者是不超过8个字符的字符串(字符为从"0"到"9"之间的数字) 交替地指定线和空白的长度, 单位为磅(points)或像素, 例如lty="44"和lty=2效果相同
lwd	控制连线宽度的数字
mar	控制图形边空的有4个值的向量c(bottom, left, top, right), 缺省值为c(5.1, 4.1, 4.1, 2.1)
mfcol	c(nr,nc)的向量, 分割绘图窗口为nr行nc列的矩阵布局, 按列次序使用各子窗口
mfrow	同上, 但是按行次序使用各子窗口
pch	控制符号的类型, 可以是1到25的整数, 也可以是""里的单个字符(见图 2.4)
ps	控制文字大小的整数, 单位为磅(points)
pty	指定绘图区域类型的字符, "s": 正方形, "m":最大利用
tck	指定轴上刻度长度的值, 单位是百分比, 以图形宽、高中最小一个作为基数; 如果tck=1则绘制grid
tcl	同上, 但以文本行高度为基数(缺省下tcl=-0.5)
xaxt	如果xaxt="n"则设置x-轴但不显示(有助于和axis(side=1, ...)联合使用)
yaxt	如果yaxt="n"则设置y-轴但不显示(有助于和axis(side=2, ...)联合使用)

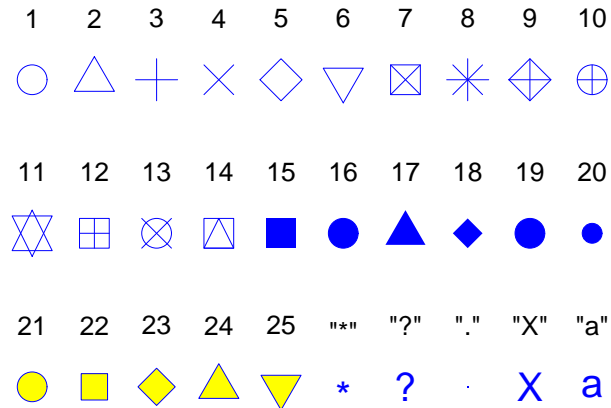


图 2.4 R (pch=1:25)的绘图符号. 用选项col="blue", bg="yellow"来产生如上的颜色, 其中背景色选项只对符号21-25有作用. 可以使用任意字符作为绘点符号(pch="*", "?", ".", ...).

2.6.4 一个实例

这一小节我们仍以R软件的内嵌数据Puromycin来说明R软件中基本的绘图方法. Puromycin的结构如下:

```
> dim(Puromycin)
[1] 23 3
> head(Puromycin)
  conc rate  state
1 0.02  76 treated
2 0.02  47 treated
3 0.06  97 treated
4 0.06 107 treated
5 0.11 123 treated
6 0.11 139 treated
```

简单的散点图(scatterplot)

对于状态(state)为treated, 画出rate关于conc的散点图, 见图2.5:

```
> PuroA <- subset(Puromycin, state == "treated")  
> plot(rate ~ conc, data = PuroA)
```

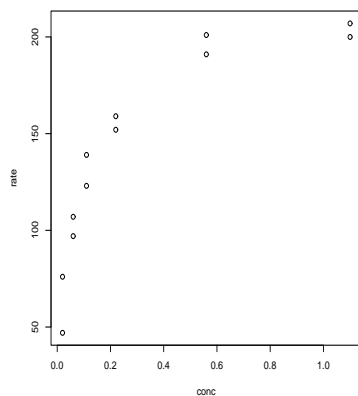


图 2.5 简单的散点图

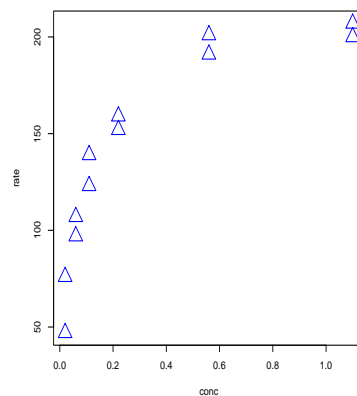


图 2.6 使用彩色符号散点图

指明所用的数据集

有三种方法指明函数plot()使用的数据集:

- 1) plot() 函数中使用data选项;
- 2) 在with() 中使用plot();

```
> with(PuroA, plot(conc, rate))
```

- 3) 使用\$ 直接指向数据与变量

```
> plot(PuroA$rate, PuroA$conc)
```

美化图形

- 1) R提供了25种不同的符号和8种不同的颜色, 浏览它们的命令是:

```
> u <- 1:25  
> plot(u ~ 1, pch = u, col = u, cex = 3)
```

- 2) 选择合适的符号及其大小与颜色. 例如图2.6是由图2.5选用绿色(选项为col=4或col="blue") 小三角形(选项为pch=2或pch="T")得到的, 大小为cex=2.5倍缺省值, 其命令为:

```
> plot(rate ~ conc, data = PuroA, pch = 2,  
       col = 4, cex = 2.5)
```

- 3) 坐标轴与标题设定. 命令

```
> plot(rate ~ conc, data = PuroA, pch = 2, col = 4,  
       cex = 2.5, xlim = c(0, 1.2), ylim = c(40, 210),  
       ylab = "Concentration",  
       xlab = "Rate", cex.lab = 2)  
> title(main = "Puromycin", cex.main = 3)
```

得到图2.7, 它做的工作有:

- 限定X轴范围为0 到1.2, Y轴范围为40 到210
- X轴标为“Rate”, Y轴标为“Concentration”
- 规定坐标轴标签大小(cex.lab=1.2)
- 增加图题

主图添线

- 1) 连接数据点. 命令

```
> library(doby) # 需要先安装  
> PuroA.mean <- summaryBy(rate ~ conc, data = PuroA,  
                          FUN = mean)  
> plot(rate ~ conc, data = PuroA, pch = 16, col = 4,  
       cex = 1.5)  
> points(mean.rate ~ conc, data = PuroA.mean, col = "cyan",  
        lwd = 10, pch = "x")  
> lines(mean.rate ~ conc, data = PuroA.mean, col = "blue")
```

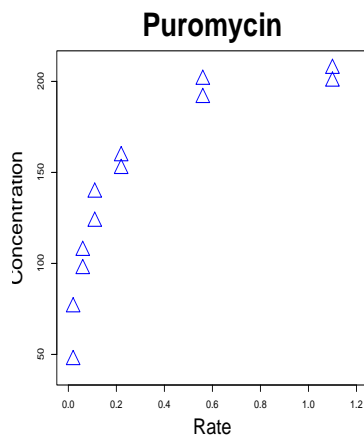



图 2.7 设定坐标轴与标题

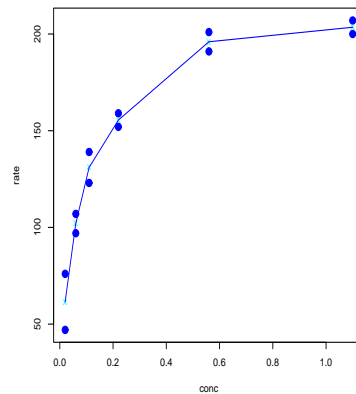


图 2.8 散点联线

得到图2.8, 它做的工作有:

- 使用doBy包的summaryBy()计算每一浓度(concentration)处的平均值
- 在每一浓度的平均值处作点
- 用直线连接这些点

2) 添加局部多项式拟合线. locfit()由局部多项式包locfit提供(需要安装). 其参数nn为光滑化参数, 用于指明曲线的光滑程度; 参数deg指明所使用的局部光滑的多项式的次数. 下面的命令给出了二条光滑曲线(见图2.9):

```
plot(rate ~ conc, data = PuroA)
smooth1 <- with(PuroA, lowess(rate ~ conc, f = 0.9))
smooth2 <- with(PuroA, lowess(rate ~ conc, f = 0.3))
lines(smooth1, col = "red")
lines(smooth2, col = "blue")
```

3) 添加多项式拟合线. 下面的命令给出了一次、二次和三次多项式拟合(见图2.10):

```
> m1 <- lm(rate ~ conc, data = PuroA)
> m2 <- lm(rate ~ conc + I(conc^2), data = PuroA)
```

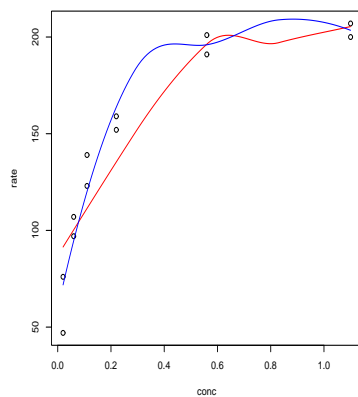


图 2.9 添加二条光滑线

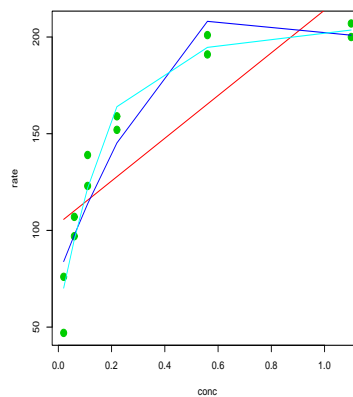


图 2.10 添加三条拟合线

```
> m3 <- lm(rate ~ conc + I(conc^2) + I(conc^3),
            data = PuroA)
> lines(fitted(m1) ~ conc, data = PuroA, col = "red")
> lines(fitted(m2) ~ conc, data = PuroA, col = "blue")
> lines(fitted(m3) ~ conc, data = PuroA, col = "cyan")
```

4) 添加参考线. 函数`abline()`可用于产生

- 回归直线: `abline(lm(...))`
- 直线: `abline(a,b)`
- 垂直线 `abline(v=a)`
- 水平线: `abline(h=b)`

命令

```
> plot(rate ~ conc, data = PuroA)
> abline(lm(rate ~ conc, data = PuroA))
> abline(a = 100, b = 105, col = "blue")
> abline(h = 200, col = "red")
> abline(v = 0.6, col = "green")
```

产生图2.11.

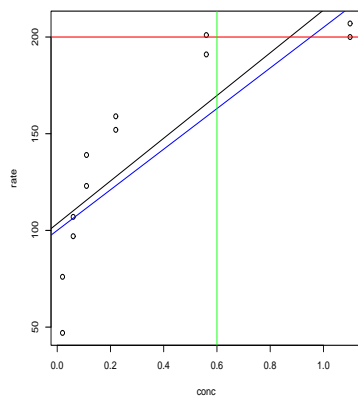


图 2.11 添加参考线

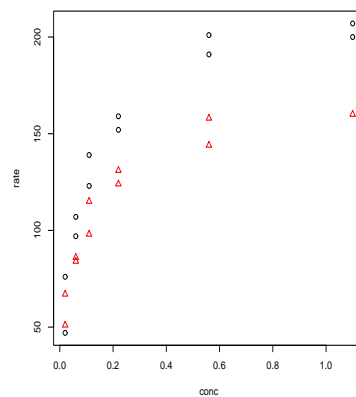


图 2.12 二个图形的叠加

图形的叠加

- 1) 两个散点图的叠加. 下面的命令将Puromycin中变量rate与conc之间的关系按state的两个值分别画出散点图. 对于“trreated”使用符号1和颜色1, 对于“untreated”使用符号2和颜色2(见图2.12), 命令为:

```
> mysymb <- c(1, 2)[Puromycin$state]
> plot(rate ~ conc, data = Puromycin, col = mysymb,
       pch = mysymb)
```

再对每一state在散点图上添加局部多项式光滑线, 产生图2.13, 命令为:

```
> PuroB <- subset(Puromycin, state == "untreated")
> smoothA <- locfit(rate ~ lp(conc, nn = 1, deg = 1),
                    data = PuroA)
> smoothB <- locfit(rate ~ lp(conc, nn = 1, deg = 1),
                    data = PuroB)
> plot(rate ~ conc, data = Puromycin, col = mysymb,
       pch = mysymb)
> lines(smoothA, lty = 1)
> lines(smoothB, lty = 3)
```

- 2) 添加图例(legend). 图2.14是在图2.13的基础上在 $(x, y) = (0.6, 100)$ 添加了图例, 其命令为:

```
> plot(rate ~ conc, data = Puromycin,
       col = c(1, 2)[state], pch = c(1, 2)[state])
> legend(x = 0.6, y = 100,
       legend = c("treated", "untreated"),
       col = c(1, 2), pch = c(1, 2), lty = c(1, 3))
```

注: 使用`locator(1)`代替`legend()`中的位置选项`x=`, `y=`可通过鼠标找到合适的位置放置图例.

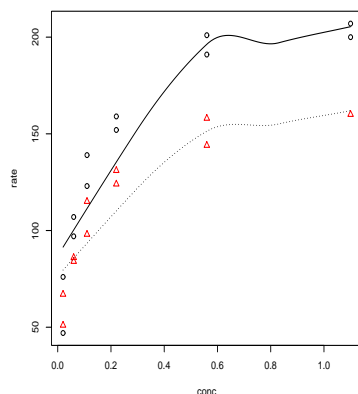


图 2.13 添加光滑线的图形叠加

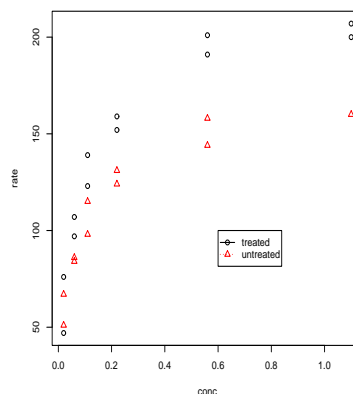


图 2.14 添加图例的图形叠加

作并列图

使用函数`par()`可以完成在同一个窗口中画多个图形, 其格式为`par(mfrow = c(m, n))`, 它表示将当前的窗口分割为 $m \times n$ 个窗口. 例如, 要在同一个窗口中作出`state`的两个值对应的两个散点图(见图2.15), 命令如下:

```
> windows(width = 7, height = 3.5)
> par(mfrow = c(1, 2))
> plot(rate ~ conc, data = PuroA)
> title("state=treated")
> plot(rate ~ conc, data = PuroB)
> title("state=untreated")
```

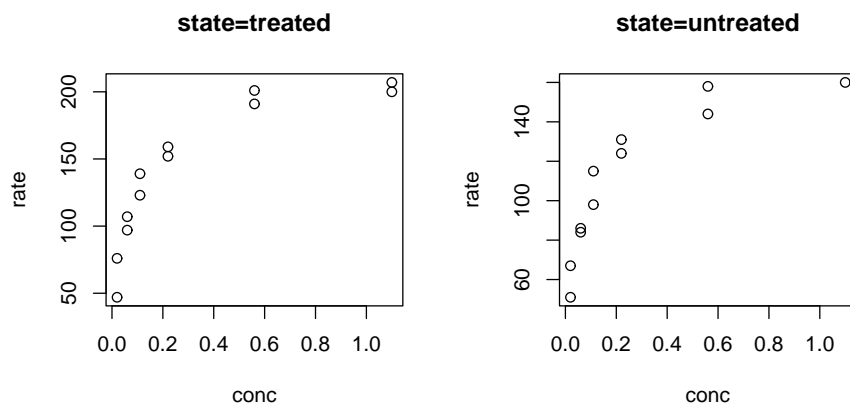


图 2.15 同一窗口的二个并列散点图.

注意:

- 要返回通常的区域中作图, 可通过命令`dev.off()`先将原来的图形关闭, 也可直接关闭图形窗口.
- 命令`par(mfrow = c(m, n))`将作图区域等分为 $m \times n$ (横向 m 行, 纵向 n 列)个窗口, 这也可以用命令`layout(matrix(1:m*n, m, n))`来实现, 但后者可以将作图区域划分为不等大小的窗口. 我们来看一下`layout()`函数中的一个例子. 命令:

```
>#-- Create a scatterplot with marginal histograms --
> x <- pmin(3, pmax(-3, stats::rnorm(50)))
> y <- pmin(3, pmax(-3, stats::rnorm(50)))
> xhist <- hist(x, breaks=seq(-3,3,0.5), plot=FALSE)
> yhist <- hist(y, breaks=seq(-3,3,0.5), plot=FALSE)
> top <- max(c(xhist$counts, yhist$counts))
> xrange <- c(-3,3); yrange <- c(-3,3)
> layout(matrix(c(2,0,1,3), 2, 2, byrow=TRUE),
           c(3,1), c(1,3), TRUE)
> layout.show(3)          # 给出作图窗口及编号
>
> par(mar=c(3,3,1,1))    # 设定边界空行数
```

```
> plot(x, y, xlim=xrange, ylim=yrange, xlab="", ylab="")
> par(mar=c(0,3,1,1))
> barplot(xhist$counts, axes=FALSE, ylim=c(0, top), space=0)
> par(mar=c(3,0,1,1))
> barplot(yhist$counts, axes=FALSE, xlim=c(0, top),
        space=0, horiz=TRUE)
```

对于向量(数据) x 和 y 同时作出它们的散点图和边际直方图. 程序的前一部分给出三个作图区域(见图2.16), 窗口1为 $3\text{cm} \times 3\text{cm}$, 用于作出 x 与 y 的散点图; 窗口2为 $3\text{cm} \times 1\text{cm}$, 用于作出 x 的散点图; 窗口3为 $1\text{cm} \times 3\text{cm}$, 用于作出 y 的散点图, 最后得到图2.17.

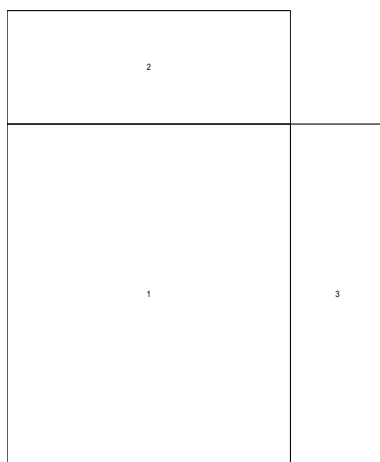


图 2.16 作图区域分割及位置

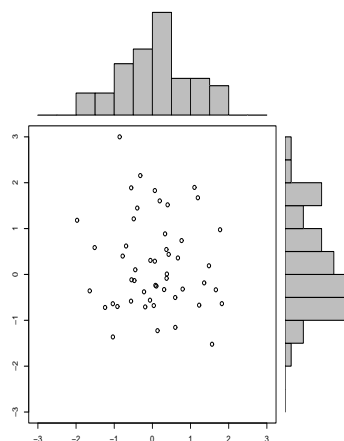


图 2.17 二维散点图及边际直方图

§2.7 R 编程

至此, 我们已经对R软件的功能有了全面的了解. 一些统计分析都是在R的对话窗口(R Console)中进行的. 但对于复杂的统计分析显然是不方便的. 下面从统计语言和编程角度来说明R编程中的一些基本技术.

2.7.1 循环和向量化

相比下拉菜单式的程序⁹, **R**的一个优势在于它可以把一系列连续的操作简单的程序化. 这一点和所有其他计算机编程语言是一致的, 但**R**有一些特性使得非专业人士也可以很简单地编写程序.

控制结构

和其他编程语言一样, **R**有一些和**C**语言(或其它语言)类似的控制结构.

- 1) 条件语句: 条件语句常用于避免除零或负数的对数等数学问题. 它有二种形式:

- `if (条件) 表达式1 else 表达式2`
- `ifelse(条件, yes, no)`

例如:

```
> if (x >= 0) sqrt(x) else NA
> ifelse(x >= 0, sqrt(x), NA)
```

- 2) 循环(loops). 它也有二种形式:

- 使用函数`for()`: `for (变量 in 向量) 表达式`
- 使用函数`while()`: `while(条件) 表达式`

两者略有区别: 若知道终止条件则用`for()`; 若无法知道运行次数, 则用`while()`. 例如, 比较下面的两种方法:

```
> for (i in 1:5) print (1:i)
> i=1
> while(i <= 5) {
  print(1:i)
  i = i+1
}
```

通常将一组命令放在大括号内. 又如, 假定我们有一个向量 x , 对于向量 x 中值为 b 的元素, 把0赋给另外一个等长度的向量 y 的对应元素, 否则赋1, 程序如下

⁹我们将在附录中介绍一个**R**下开发的菜单式软件: **R Commander**.

```
> y <- numeric(length(x)) #创建一个x等长的向量y
> for (i in 1:length(x)){
  if (x[i] == b)
    y[i] <- 0
  else
    y[i] <- 1
}
```

向量化(vectorization)

在R中, 很多情况下循环和控制结构可以通过向量化避免(简化): 向量化使得循环隐含在表达式中. 比如, 条件语句也可以用逻辑索引向量代替. 前面的例子可以改写为:

```
> y[x == b] <- 0
> y[x != b] <- 1
```

在实际编程时, 如果能将一组命令向量化, 则应尽量避免循环, 原因在于

- 代码更简洁
- C是一种编译语言, 其效率是很高的; R则是一种解释语言. 在计算时, 通常C要比R快100倍.
- 在R中使用向量化, R会立即调用C进行运算, 因而大大提高计算的效率.

2.7.2 用R写程序

一般情况下, 一个R程序以ASCII 格式保存, 扩展名为'.R'. 如果一个工作要重复好多次, 用R程序是一个不错的选择. 考虑这样的例子: 我们想对三种不同的鸟绘制一样的图, 而且数据在三个不同的文件中. 我们将一步一步的演示二种不同的方式, 看R是如何完成这个简单的过程.

首先, 我们凭直觉连续键入一系列命令, 而且预先分割图形界面:

```
> layout(matrix(1:3, 3, 1))           #分割图形界面
> data <- read.table("Swal.dat")       #读入数据
> plot(data$V1, data$V2, type="l")
```



```
> title("swallow")                #增加标题
> data <- read.table("Wren.dat")
> plot(data$V1, data$V2, type="l")
> title("wren")
> data <- read.table("Dunn.dat")
> plot(data$V1, data$V2, type="l")
> title("dunnock")
```

我们看到一些命令多次执行, 因此它们可以放在一起, 在执行的时候仅仅修改一些参数. 这里的策略是把参数放到一个字符型的向量中去, 然后用下标去访问这些不同的值. 修改后的程序如下:

```
> layout(matrix(1:3, 3, 1))        # 分割图形界面
> species <- c("swallow", "wren", "dunnock")
> file <- c("Swal.dat", "Wren.dat", "Dunn.dat")
> for(i in 1:length(species)) {
  data <- read.table(file[i])        # 读入数据
  plot(data$V1, data$V2, type="l")
  title(species[i])                 # 增加标题
}
```

如果程序保存在文件Mybirds.R 中, 可以通过键入如下命令执行:

```
> source("Mybirds.R")
```

注意: 和所有以文件作为输入对象的函数一样, 如果该文件不在当前工作目录下, 用户需要提供该文件的绝对路径.

2.7.3 编写你自己的函数

大多数R的工作是通过函数来实现的, 而且这些函数的输入参数都放在一个括弧里面. 用户可以编写自己的函数, 并且这些函数和R里面的其它函数有一样的特性.

函数是一系列语句的组合, 形式为:

函数定义的基本形式

```
变量名 = function( 变量列表 ) 函数体
```

编写自己的函数可以让你有效、灵活、合理地使用R. 我们再次使用前面读数据并且画图的例子. 如果我们想在其它情况下进行这样的操作, 写一个函数是一个不错的想法:

```
> myfun <- function(S, F) {
  data <- read.table(F)
  plot(data$V1, data$V2, type="l")
  title(S)
}
```

执行时, 这个函数必须载入内存. 一旦函数载入后, 我们就可以键入一条命令以读入数据和画出我们想要的图. 因此, 现在我们的程序有第三个实现的版本了:

```
> layout(matrix(1:3, 3, 1))
> myfun("swallow", "Swal.dat")
> myfun("wren", "Wrenn.dat")
> myfun("dunnock", "Dunn.dat")
```

我们还可以用`sapply()`¹⁰实现程序的第四个版本:

```
> layout(matrix(1:3, 3, 1))
> species <- c("swallow", "wren", "dunnock")
> file <- c("Swal.dat", "Wren.dat", "Dunn.dat")
> sapply(species, myfun, file)
```

函数的调用与其参数的位置与名字(又称为标签参数)有关, 假定函数`foo1()`有三个参数, 其定义为:

```
> foo1 <- function(arg1, arg2, arg3) {...}
```

¹⁰对于向量或列表`X`和作用它们的函数“`Fun`”, 在R中可使用命令`lapply(X, Fun)`和`sapply(X, Fun)`, 两者的差异仅在于: 前者返回与`X`长度相等的一个列表, 后者返回一个向量或矩阵. 两者本质上相同, 后者只是为前者的友好形式.

则计算函数`foo(x,y,z)`在 (u,v,w) 处的值, 可以采用下面两种方法中的一种:

```
> foo1(u, v, w)           # 按位置调用函数
> foo1(arg3=w, arg2=v, arg1=u) # 按名字调用
```

R函数的另外一个特性是函数调用可以采用定义时的默认设置. 例如函数`foo2()`也有三个参数, 其定义为:

```
> foo2 <- function(arg1, arg2 = 5, arg3 = FALSE) {...}
```

则下面的三种命令等价

```
> foo2(x)}
> foo2(x, 5, FALSE)
> foo2(x, arg3 = FALSE)
```

使用一个函数的默认设置非常有用, 特别在使用标签参数的时候, 例如

```
> foo2(x, arg3 = TRUE)
```

仅仅改变一个默认设置.

在结束本章前, 我们来看另外一个例子. 尽管这个例子不是纯粹的统计学例子, 但是它很好地展示了**R**语言的灵活性. 假定我们想研究一个非线性模型的行为: 这个模型 (Ricker 模型) 的定义如下:

$$N_{t+1} = N_t \exp \left[r \left(1 - \frac{N_t}{K} \right) \right]$$

这个模型广泛地用于种群动态变化的研究, 特别是鱼类的种群变化. 我们想用 一个函数去模拟这个模型关于增长率 r 和初始群体大小 N_0 的变化情况(承载能力 K 常常设定为1且以这个值作为默认值); 结果将以种群大小相对时间的图表示. 我们还将设定一个可选项允许用户只显示最后若干步中种群大小(默认所有结果都会被绘制出来). 下面的函数就是Ricker模型的数值模拟.

```
> ricker <- function(nzero, r, K=1, time=100, from=0, to=time) {
  N <- numeric(time+1)
  N[1] <- nzero
```

```
for (i in 1:time) N[i+1] <- N[i]*exp(r*(1 - N[i]/K))
Time <- 0:time
plot(Time, N, type="l", xlim=c(from, to))
}
```

你可以试一试下面的代码:

```
> layout(matrix(1:3, 3, 1))
> ricker(0.1, 1); title("r = 1")
> ricker(0.1, 2); title("r = 2")
> ricker(0.1, 3); title("r = 3")
```

2.7.4 养成良好的编程习惯

为了他人, 更为你本人! 你的程序应该具有

- 可读性(readability)
- 可理解性(understandability)

为此你应该养成四个良好的习惯:

习惯之一: 采用结构化、模块化编程;

习惯之二: 增加注释(Commenting), **R**中使用# 作为注释语句的开始;

习惯之三: 使用意义明确的名字给变量命名, 切忌使用人或宠物的名字;

习惯之四: 行前自动缩进(Indentation), 在此推荐使用软件WinEdt, 现在有针对**R**的平台: RWinEdt. 见附录A的具体介绍.

第二章习题

2.1 用函数`rep()`构造一个向量 x , 它由3个3, 4个2, 5个1构成.

2.2 由 $1, 2, \dots, 16$ 构成二个方阵, 其中矩阵 A 按列输入, 矩阵 B 按行输入, 并计算:

1) $C = A + B$;

2) $D = AB$;

3) $E = (e_{ij})_{n \times n}$;

4) 去除 A 和第3行, B 的第3列, 重新计算上面的矩阵 E .

2.3 函数`solve()`有二个作用: `solve(A,b)`可用于求解线性方程组 $\mathbf{A}x = b$, `solve(A)`可用于求矩阵 A 的逆. 设

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 10 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

用二种方法编程求方程组 $\mathbf{A}x = b$ 的解.

2.4 设 x 与 y 表示 n 维的向量, 则`x%*%y`或`crossprod(x,y)`用于求它们的内积, 即`t(x)%*%y`; 而`x%o%y`或`outer(x, y)`用于求它们的外积(叉积), 即`x%*%t(y)`, 其中`t()`表示矩阵或向量的转置. 设 $x = (1, 2, 3, 4, 5)$, $y = (2, 4, 6, 8, 10)$. 用三种不同的方法求它们的内积与外积.

2.5 编写一个用二分法求非线性方程根的函数, 并求方程

$$x^3 - x - 1 = 0$$

在区间 $[1, 2]$ 内的根, 精度要求 $\epsilon = 10^{-5}$.

2.6 自己编写一个函数, 求数据 $y = (y_1, y_2, \dots, y_n)$ 的均值、标准差、偏度与峰度.

2.7 有10名学生的身高与体重数据如表2.7所示.

1) 用数据框的形式读入数据;

2) 将数据表2.7写成一个纯文本的文件, 并用函数`read.table()`读取该文件中的数据;

表 2.7 学生身高与体重数据

序号	性别	年龄	身高/cm	体重/kg
1	F	14	156	42.3
2	F	16	158	45.0
3	F	15	161	48.5
4	F	17	156	51.5
5	F	15	153	44.6
6	M	14	162	48.8
7	M	16	157	46.7
8	M	14	159	49.9
9	M	15	163	50.2
10	M	16	165	53.7

3) 用函数`write.csv()`写成一个能用Excel打开的文件, 测试是否成功.

第三章 概率与分布

本章概要

- ◇ 随机抽样的实现
- ◇ 常用的概率分布及其数字特征
- ◇ **R**中内嵌的分布

§3.1 随机抽样

众所周知, 概率论早期研究的是游戏或赌博等随机现象中有关的概率问题. 这些现象在**R**中可以通过函数`sample()`来实现.

1) 等可能的不放回的随机抽样:

```
> sample(x, n)
```

其中`x`为要抽取的向量, `n`为样本容量. 例如从52张扑克牌中抽取4张对应的**R**命令为:

```
> sample(1:52, 4)
[1]  3 16 17 15
```

2) 等可能的有放回的随机抽样:

```
> sample(x, n, replace=TRUE)
```

其中选项`replace=TRUE`表示抽样是有放回的, 此选项省略或为`replace=FALSE`表示抽样是不放回的. 例如抛一枚均匀的硬币10次在**R**中可表示为:

```
> sample(c("H", "T"), 10, replace=T)
[1] "H" "T" "T" "H" "H" "T" "T" "H" "H" "H"
```

掷一棵骰子10次可表示为:

```
> sample(1:6, 10, replace=T)
[1] 4 3 4 5 4 6 2 6 3 4
```

3) 不等可能的随机抽样:

```
> sample(x, n, replace=TRUE, prob=y)
```

其中选项`prob=y`用于指定`x`中元素出现的概率, 向量`y`与`x`等长度. 例如一名外科医生做手术成功的概率为0.90, 那么他做10次手术在`R`中可以表示为:

```
> sample(c("成功", "失败"), 10, replace=T, prob=c(0.9,0.1))
```

若以1表示成功, 0表示失败, 则上述命令可变为:

```
> sample(c(1,0), 10, replace=T, prob=c(0.9,0.1))
[1] 1 1 1 0 1 1 1 1 1 1
```

§3.2 排列组合与概率的计算

我们仍以扑克牌为例加以说明.

例 3.2.1 从一副完全打乱的52张扑克中取4张, 求以下事件的概率:

- 1) 抽取的4张依次为红心A, 方块A, 黑桃A和梅花A的概率;
- 2) 抽取的4张为红心A, 方块A, 黑桃A和梅花A的概率.

解

- 1) 抽取的4张是有次序的, 因此使用排列来求解. 所求的事件(记为A)概率为

$$P(A) = \frac{1}{52 \times 51 \times 50 \times 49}.$$

在`R`中计算得到


```
> 1/prod(52:49)
[1] 1.539077e-07
```

- 2) 抽取的4张是没有次序的, 因此使用组合数来求解. 所求的事件(记为B)概率为

$$P(B) = \frac{1}{\binom{52}{4}},$$

其中 $\binom{n}{m} = \frac{n!}{m!(n-m)!}$. 在R中计算得到

```
> 1/choose(52,4)
[1] 3.693785e-06
```

■

§3.3 概率分布

概率论与数理统计是研究随机现象统计规律性的一门学科. 对于一个具体的问题, 通常归结为对一个随机变量或随机向量(X)的取值及其取值概率的研究, 即对于事件 $P(X \leq x)$ 的研究. 这就是随机变量的累积分布函数(CDF), 记为 $F(x)$. 因此随机变量统计规律可以用累积分布函数来刻画. 对于离散型随机变量(取值为有限或可列无限), 其统计规律通常转化为对分布律 $f(x) = P(X = x)$ 的研究, 它与分布函数的关系为 $F(x) = \sum_{t \leq x} P(X = t)$; 而对于连续型随机变量(取值充满整个区间), 其统计规律通常转化为对概率密度函数 $f(x)$ 的研究, 它与分布函数的关系为 $F(x) = \int_{-\infty}^x f(x)dx$. 下面我们分离散与连续二种情况分别介绍它们的分布律或密度函数, 在此我们不加区分地使用 $f(x)$.

3.3.1 离散分布的分布律

- 1) 贝努里分布: `binom(1, p)`

- 意义: 一试验中有二个事件: 成功(记为1)与失败(记为0), 出现的概率是分别为 p 和 $1 - p$, 则一次试验(称为贝努里试验)成功的次数服从一个参数为 p 的贝努里分布.
- 分布律:

$$f(x|p) = p^x(1-p)^{1-x}, \quad x = 0, 1 \quad (0 < p < 1).$$

- 数字特征: $E(X) = p, \text{Var}(X) = p(1 - p)$.

2) 二项分布: $\text{binom}(n, p)$

- 意义: 贝努里试验独立地重复 n 次, 则试验成功的次数服从一个参数为 (n, p) 的二项分布.
- 分布律:

$$f(x|n, p) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, \dots, n.$$

- 数字特征: $E(X) = np, \text{Var}(X) = np(1 - p)$.
- 特例: $n = 1$ 时分布为贝努里分布.

3) 多项分布: $\text{multinom}(n, p_1, \dots, p_k)$

- 意义: 一试验中有 k 个事件 $A_i, i = 1, 2, \dots, k$, 且 $P(A_i) = p_i$ ($0 < p_i < 1, \sum_{i=1}^k p_i = 1$). 将此试验独立地重复 n 次, 则事件 A_1, A_2, \dots, A_k 出现的次数服从一个参数为 (n, \mathbf{p}) 的多项分布, 其中 $\mathbf{p} = (p_1, p_2, \dots, p_k)$.
- 分布律:

$$f(x_1, \dots, x_k | n, \mathbf{p}) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}, \quad 0 \leq x_i \leq n, \sum_{i=1}^k x_i = n.$$

- 数字特征: $E(X_i) = np, \text{Var}(X_i) = np(1 - p), \text{Cov}(X_i, X_j) = -np_i p_j$.
- 特例: $k = 2$ 时分布为二项分布.

4) 负二项分布: $\text{nbinom}(k, p)$

- 意义: 贝努里试验独立地重复进行, 一直到出现 k 次成功时停止试验, 则试验失败的次数服从一个参数为 (k, p) 的负二项分布.
- 分布律:

$$f(x|k, p) = \frac{\Gamma(k+x)}{\Gamma(k)\Gamma(x)} p^k (1-p)^x, \quad x = 0, 1, \dots$$

- 数字特征: $E(X) = \frac{k(1-p)}{p}, \text{Var}(X) = \frac{k(1-p)}{p^2}$.

- 特例: $k = 1$ 时的分布为几何分布.

5) 几何分布: $\text{geom}(p)$

- 意义: 努里试验独立地重复进行, 一直到出现有成功出现时停止试验, 则试验失败的次数服从一个参数为 p 的几何分布.
- 分布律:

$$f(x|p) = p(1-p)^x, \quad x = 0, 1, 2, \dots$$

- 数字特征: $E(X) = \frac{(1-p)}{p}, \text{Var}(X) = \frac{(1-p)}{p^2}.$

6) 超几何分布: $\text{hyper}(N, M, n)$

- 意义: 从装有 N 个白球和 M 个黑球的罐子中不放回地取出 k ($\leq N + M$) 个球, 则其中的白球数服从超几何分布.
- 分布律:

$$f(x|N, M, k) = \frac{\binom{N}{x} \binom{M}{k-x}}{\binom{N+M}{k}}, \quad x = 0, 1, 2, \dots, \min\{N, k\}.$$

- 数字特征: $E(X) = \frac{(kN)}{N+M}, \text{Var}(X) = \left(\frac{N+M-k}{N+M-1}\right) \frac{kN}{N+M} \left(1 - \frac{N}{N+M}\right).$

7) 泊松分布: $\text{pois}(\lambda)$

- 意义: 单位时间, 单位长度, 单位面积, 单位体积中发生某一事件的次数常可以用泊松(Poisson)分布来刻画, 例如某段高速公路上一年内的交通事故数和某办公室一天中收到的电话数可以认为近似服从泊松分布.
- 分布律:

$$f(x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 1, 2, \dots$$

- 数字特征: $E(X) = \lambda, \text{Var}(X) = \lambda.$

3.3.2 连续分布的密度函数

1) 贝塔分布: $\text{Beta}(a, b)$

- 意义: 在贝叶斯分析中, 贝塔分布常作为二项分布参数的共轭先验分布.

- 密度函数:

$$f(x|a, b) = \frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1}, \quad 0 < x < 1 \quad (a, b > 0).$$

- 数字特征: $E(X) = \frac{a}{a+b}$, $\text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$.
- 特例: $a = 1, b = 1$ 时的分布为 $[0, 1]$ 上的均匀分布.

2) 均匀分布: $\text{unif}(a, b)$

- 意义: 区间 $[a, b]$ 上随机投点对应的坐标服从 $[a, b]$ 上的均匀分布.
- 密度函数:

$$f(x|a, b) = \frac{1}{b-a}, \quad a \leq x \leq b.$$

- 数字特征: $E(X) = \frac{a+b}{2}$, $\text{Var}(X) = \frac{b^2-a^2}{12}$.

3) 柯西分布: $\text{cauchy}(a, b)$

- 意义: 柯西分布(又称为Lorentz分布)用于描述共振行为. 以一随机的角度投向 X 轴的水平距离服从柯西分布.
- 密度函数:

$$f(x|a, b) = \frac{1}{\pi b \left[1 + \left(\frac{x-a}{b} \right)^2 \right]}, \quad 0 < x < 1 \quad (a, b > 0).$$

- 数字特征: 均值与方差不存在.

4) 威布尔分布: $\text{weibull}(a, b)$

- 意义: 最为常用的寿命分布, 用来刻画滚珠轴承、电子元器件等产品的寿命.
- 密度函数:

$$f(x|a, b) = abx^{b-1}e^{-ax^b}, \quad x > 0 \quad (a, b > 0).$$

- 数字特征: $E(X) = \frac{\Gamma(1 + \frac{1}{b})}{a^{1/b}}$,
 $\text{Var}(X) = \frac{\Gamma(1 + \frac{2}{b})}{a^{2/b}} - \frac{\{\Gamma(1 + \frac{1}{b})\}^2}{a^{2/b}}.$
- 特例: $b = 1$ 时的分布为指数分布.

5) 指数分布: $\text{exp}(\lambda)$

- 意义: 泊松过程的等待时间服从指数分布. 形状参数 $b = 1$ 的Weibull分布为指数分布.

- 密度函数:

$$f(x|\lambda) = \lambda e^{-\lambda x}, \quad x > 0 \quad (\lambda > 0).$$

- 数字特征: $E(X) = \frac{1}{\lambda}, \text{Var}(X) = \frac{1}{\lambda^2}.$

6) 瑞利(Rayleigh)分布: $\text{rayl}(b)$

- 意义: 瑞利(Rayleigh)分布为Weibull分布的又一个特例: 它是参数为 $(1/(2b^2), 2)$ 的Weibull分布.

- 密度函数:

$$f(x|b) = \frac{x}{b^2} \exp\left(-\frac{x^2}{2b^2}\right).$$

- 数字特征: $E(X) = \sqrt{\frac{\pi}{2}}b, \text{Var}(X) = \frac{4-\pi}{2}b^2.$

7) 正态分布/高斯分布: $\text{norm}(\mu, \sigma^2)$

- 意义: 高斯分布是概率论与数理统计中最重要的一个分布. 中心极限定理表明, 一个变量如果是由大量微小的、独立的随机因素的叠加结果, 那么这个变量一定是正态变量. 因此许多随机变量可以用高斯分布表述或近似描述..

- 密度函数:

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty, \\ (-\infty < \mu < \infty, \sigma > 0)$$

- 数字特征: $E(X) = \mu, \text{Var}(X) = \sigma^2.$

8) 对数正态分布: $\text{lnorm}(\mu, \sigma^2)$

- 意义: $\ln(X)$ 服从参数为 (μ, σ^2) 的正态分布, 则 X 服从参数为 (μ, σ^2) 的对数正态分布.

- 密度函数:

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, \quad x > 0 \\ (-\infty < \mu < \infty, \sigma > 0)$$

- 数字特征: $E(X) = \exp\left\{\mu + \frac{1}{2}\sigma^2\right\}$, $\text{Var}(X) = e^{\sigma^2}(e^{\sigma^2} - 1)e^{2\mu}$.

9) 逆正态分布: **inorm**(μ, λ)

- 意义: 正态随机变量的倒数服从的分布.
- 密度函数:

$$f(x|\mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp -\frac{\lambda(x - \mu)}{2\mu^2 x} \quad (-\infty < \mu < \infty, \lambda > 0)$$

- 数字特征: $E(X) = \mu$, $\text{Var}(X) = \frac{\mu^3}{\lambda}$.

10) 伽玛分布: **gamma**(a, b)

- 意义: k 个相互独立的参数为 $1/b$ 的指数分布的和服从参数为 (k, b) 的伽玛分布.
- 密度函数:

$$f(x|a, b) = \frac{1}{\Gamma(a)b^a} x^{a-1} e^{-x/b}, \quad x > 0 \quad (a > 0, b > 0).$$

- 数字特征: $E(X) = ab$, $\text{Var}(X) = ab^2$.
- 特例: $a = 1$ 时的分布为指数分布; $a = \frac{n}{2}, b = 2$ 时的分布为卡方分布.

11) 逆伽玛分布: **igamma**(a, b)

- 意义: 伽玛分布随机变量的倒数服从逆伽玛分布.
- 密度函数:

$$f(x|a, b) = \frac{1}{\Gamma(a)b^a} x^{-(a+1)} e^{-1/(bx)}, \quad x > 0 \quad (a > 0, b > 0).$$

- 数字特征: $E(X) = \frac{1}{b(a-1)} (a > 1)$, $\text{Var}(X) = \frac{1}{b^2(a-1)^2(a-2)} (a > 2)$.
- $a = \frac{n}{2}, b = 2$ 的分布为逆卡方分布.

12) 卡方(χ^2)分布: **chisq**(n)

- 意义: n 个独立正态随机变量的平方和服从自由度为 n 的卡方分布.
- 密度函数:

$$f(x|n) = \frac{x^{n/2-1} e^{-x/2}}{2^{n/2} \Gamma(n/2)}, \quad x > 0.$$

- 数字特征: $E(X) = n, \text{Var}(X) = 2n \quad (n > 2)$.

13) 逆卡方分布: $\text{ichisq}(n)$

- 意义: 卡方分布随机变量的倒数服从逆卡方分布.
- 密度函数:

$$f(x|n) = \frac{x^{-(n/2+1)} e^{-1/2x}}{2^{n/2} \Gamma(n/2)}, \quad x > 0.$$

- 数字特征: $E(X) = \frac{1}{n-2} \quad (n > 2), \text{Var}(X) = \frac{2}{(n-2)^2(n-4)} \quad (n > 4)$.

14) t 分布: $t(n)$

- 意义: 随机变量 X 与 Y 独立, X 服从标准正态分布, Y 服从自由度为 n 卡方分布, 则 $T = \frac{X}{\sqrt{Y/n}}$ 服从自由度为 n 的 t 分布.

- 密度函数:

$$f(x|n) = \frac{\left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}}{\sqrt{n} B\left(\frac{1}{2}, \frac{n}{2}\right)}.$$

- 数字特征: $E(X) = 0, \text{Var}(X) = \frac{n}{n-2} \quad (n > 2)$.

15) F 分布: $f(n, m)$

- 意义: 随机变量 X 与 Y 独立, X 服从自由度为 n 卡方分布, Y 服从自由度为 m 卡方分布, 则 $T = \frac{X/n}{Y/m}$ 服从自由度为 (n, m) 的 F 分布.

- 密度函数:

$$f(x|n, m) = \frac{\left(\frac{n}{m}\right)^{n/2} x^{n-2}/2}{B\left(\frac{n}{2}, \frac{m}{2}\right)} \left(1 + \frac{n}{m}x\right)^{-(n+m)/2}.$$

- 数字特征: $E(X) = \frac{m}{m-2} \quad (m > 2), \text{Var}(X) = \frac{2m^2(n+m-2)}{n(m+2)} \quad (n > 2)$.

16) logistic分布: $\text{logis}(a, b)$

- 意义: 生态学中的增长模型常用logistic分布来刻画, 它也常用于logistic回归中.

- 密度函数:

$$f(x|a, b) = \left[1 + e^{-(x-a)/b}\right]^{-1}.$$

- 数字特征: $E(X) = a, \text{Var}(X) = \frac{\pi^2}{3}b^2$.

17) Dirichlet分布: $\text{Dirichlet}(\alpha_1, \dots, \alpha_k)$

- 意义: 在贝叶斯分析中可作为多项分布参数的共轭分布. Dirichlet分布的密度函数表示在已知 k 个竞争事件已经出现了 $\alpha_i - 1$ 次条件下, 它们出现的概率为 $x_i, i = 1, 2, \dots, k$ 的信念.
- 密度函数:

$$f(x_1, \dots, x_k | \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^k x_i^{\alpha_i-1}, \quad x_i > 0, \sum_{i=1}^k x_i = 1 \quad (\alpha_i > 0),$$

$$\text{其中 } B(\alpha) = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^k \alpha_i)}.$$

- 数字特征: $E(X) = \frac{\alpha_i}{\alpha_0}, \quad \text{Var}(X) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)},$
 $\text{Cov}(X_i, X_j) = -\frac{\alpha_0 \alpha_i}{\alpha_0^2(\alpha_0 + 1)},$ 其中 $\alpha_0 = \sum_{i=1}^k \alpha_i$.
- $k = 2$ 为贝塔分布.

18) Pareto分布: $\text{pd}(a, b)$

- 意义: 财富的分配的规则(称为Pareto规则)是大部分的财富(80%)被少数人(20%)的人拥有, 这可以较好地用Pareto分布来刻画.
- 密度函数:

$$f(x|a, b) = \frac{b}{a} \left(\frac{a}{x}\right)^{b+1}, \quad x > a \quad (b > 0).$$

- 数字特征: $E(X) = \frac{a}{b-1} \quad (b > 1), \text{Var}(X) = \frac{a^2 b}{(b-1)^2(b-2)} \quad (b > 2).$

19) 非中心分布. 与前面卡方分布、t分布和F分布相对应还有三个非中心的分布:

- 非中心的卡方分布 — $\text{chisq}(n, \mu)$: n 个独立正态随机变量 $N(\mu_i, \sigma^2), i = 1, 2, \dots, n$ 的平方和服从自由度为 n 、非中心参数为 $\mu = \frac{\mu_1^2 + \mu_2^2 + \dots + \mu_n^2}{\sigma^2}$ 的卡方分布.
- 非中心的t分布 — $\text{t}(n, \mu)$: 随机变量 X 与 Y 独立, X 服从标准正态分布, Y 服从自由度为 n 卡方分布, 则 $T = \frac{X+\mu}{\sqrt{Y/n}}$ 服从自由度为 n 、非中心参数为 μ 的 t 分布.

- 非中心的F分布 — $F(n, m, \mu)$: 随机变量 X 与 Y 独立, X 服从自由度为 n 、非中心参数为 μ 的非中心卡方分布, Y 服从自由度为 m 卡方分布, 则 $T = \frac{X/n}{Y/m}$ 服从自由度为 (n, m) 、非中心参数为 μ 的F分布.

若无特别申明, 通常所说的卡方分布、t分布和F分布都是中心的卡方分布、t分布和F分布.

§3.4 R中内嵌的分布

R提供了四类有关统计分布的函数: 密度函数、(累积)分布函数、分位数函数、随机数函数. 它们都与分布的英文名称(或者其缩写)相对应. 下表按英文字母顺序列出R中提供了18个分布的英文名称、R中的名称和函数中的选项:

分布名称	R名称	选项
beta	beta	shape1, shape2
binomial	binom	size, prob
Cauchy	cauchy	location=0, scale=1
chi-squared (χ^2)	chisq	df, ncp
exponential	exp	rate
Fisher-Snedecor (F)	f	df1, df2, ncp
gamma	gamma	shape, scale=1
geometric	geom	prob
hypergeometric	hyper	m, n, k
lognormal	lnorm	meanlog=0, sdlog=1
logistic	logis	location=0, scale=1
multinomial	multinom	size, prob
normal	norm	mean=0, sd=1
negative binomial	nbinom	size, prob
Poisson	pois	lambda
Student's (t)	t	df
uniform	unif	min=0, max=1
Weibull	weibull	shape, scale=1
Wilcoxon's statistics	wilcox	m, n
	signrank	n

对于所给的分布名称, 加前缀“d”(代表密度函数, `density`)就得到 \mathbf{R} 的密度函数(对于离散分布, 指分布律); 加前缀“p”(代表分布函数或概率, CDF)就得到 \mathbf{R} 的分布函数; 加前缀“q”(代表分位函数, `quantile`)就得到 \mathbf{R} 的分位数函数; 加前缀“r”(代表随机模拟, `random`)就得到 \mathbf{R} 的随机数发生函数. 而且这四类函数的第一个参数是有规律的: 形为`dfunc`的函数为 x , `pfunc`的函数为 q , `qfunc`的函数为 p , `rfunc`的函数为 n (但`rhyp`和`rwilcox`是特例, 他们的第一个参数为`nm`). 目前为止, 非中心参数(non-centrality parameter)仅对CDF和少数其它几个函数有效, 细节请参考在线帮助.

若 \mathbf{R} 中分布的函数名为`func`, 则四类函数的调用格式为:

- 1) 概率密度函数: `dfunc(x, p1, p2, ...)`, x 为数值向量;
- 2) (累积)分布函数: `pfunc(q, p1, p2, ...)`, q 为数值向量;
- 3) 分位数函数: `qfunc(p, p1, p2, ...)`, p 为由概率构成的向量;
- 4) 随机数函数: `rfunc(n, p1, p2, ...)`, n 为生成数据的个数

其中`p1, p2, ...`是分布的参数值. 上面的表格中有具体数值的是这些参数在空缺时对应的缺省值.

所有`pfunc`和`qfunc`的函数都具有逻辑参数`lower.tail`和`log.p`, 而所有的`dfunc`函数都有参数`log`. 此外, 对于来自正态分布, 具有学生化样本区间的分布还有`ptukey`和`qtukey`这样的函数.

最后通过二个例子简单说明一下它们的作用:

- 1) 查找分布的分位数, 用于计算假设检验中分布的临界值或置信区间的置信限. 例如, 显著性水平为5%的正态分布的双侧临界值是:

```
> qnorm(0.025)
[1] -1.959964
> qnorm(0.975)
[1] 1.959964
```

- 2) 计算假设检验的 p 值. 比如自由度 $df = 1$ 的 $\chi^2 = 3.84$ 时的 χ^2 检验的 p 值为

```
> 1 - pchisq(3.84, 1)
[1] 0.05004352
```

而容量为14的双边t检验的p值为

```
> 2*pt(-2.43, df = 13)
[1] 0.0303309
```

这些函数将在以后的章节中发挥极大的作用.

§3.5 应用：中心极限定理

3.5.1 中心极限定理

正态分布在概率统计中起着至关重要的作用，其中的一个原因是当独立观察(试验)的样本容量 n 足够大时，那么所观察的随机变量 X_1, X_2, \dots, X_n 的和近似服从正态分布(假定 $E(X_i) = \mu, \text{Var}(X_i) = \sigma^2$ 存在)，即

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \sim N(0, 1) \quad (n \rightarrow \infty)$$

或

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad (n \rightarrow \infty)$$

3.5.2 渐近正态性的图形检验

下面的函数给出了从图形上考查一个由(**R**中已经提供的或自己定义的)已知分布产生的容量为 n 的样本(可以为向量)经标准化变换后趋于标准正态分布的近似程度.

```
limite.central( )的定义

limite.central <- function (r=runif, distpar=c(0,1), m=.5,
                           s=1/sqrt(12),
                           n=c(1,3,10,30), N=1000) {
  for (i in n) {
    if (length(distpar)==2){
      x <- matrix(r(i*N, distpar[1],distpar[2]),nc=i)
    }
    else {
      x <- matrix(r(i*N, distpar), nc=i)
    }
  }
}
```

```
x <- (apply(x, 1, sum) - i*m)/(sqrt(i)*s)
hist(x,col='light blue',probability=T,main=paste("n=",i),
      ylim=c(0,max(.4, density(x)$y)))
lines(density(x), col='red', lwd=3)
curve(dnorm(x), col='blue', lwd=3, lty=3, add=T)
if( N>100 ) {
  rug(sample(x,100))
}
else {
  rug(x)
}
}
```

此函数的缺省值为:

- 1) 分布为 $[0, 1]$ 上的均匀分布, 否则用选项**r**=声明;
- 2) 分布的均值为0.5, 否则用选项**m**=声明;
- 3) 分布的标准差为 $1/\sqrt{12}$, 否则用选项**s**=声明;
- 4) 样本容量有4个: 1, 3, 10, 30, 否则用选项**n**=声明;
- 5) 重复次数为1000, 否则用选项**N**=声明.

对于程序作一简单说明:

- 1) **hist(x, ...)**用于作出 x 的直方图;
- 2) **lines(density(x), ...)**计算 x 的核密度估计值(窗宽为**bw**=1), 并连接成线;
- 3) **curve(dnorm(x), ...)**计算 x 处标准正态分布的密度函数值, 并连接成线;
- 4) **rug(x)**在横坐标处用小的竖线画出 x 出现的位置.

有关的其它参数, 参见第二章的说明或通过**R**关于这些函数的帮助. 如果将程序中的 x 改为样本的标准化值, 就可检验一般样本的渐近正态性.

3.5.3 举例

二项分布: $b(10, 0.1)$

```
op <- par(mfrow=c(2,2))
limite.central(rbinom, distpar=c(10 ,0.1), m=1, s=0.9)
par(op)
```

得到图3.1.

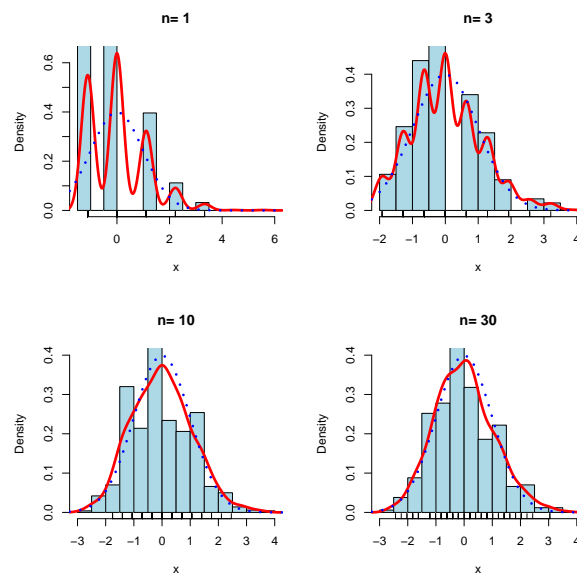


图 3.1 二项分布的渐近正态性.

泊松分布: $\text{pios}(1)$

```
op <- par(mfrow=c(2,2))
limite.central(rpois, distpar=1, m=1, s=1, n=c(3, 10, 30 ,50))
par(op)
```

得到图3.2.

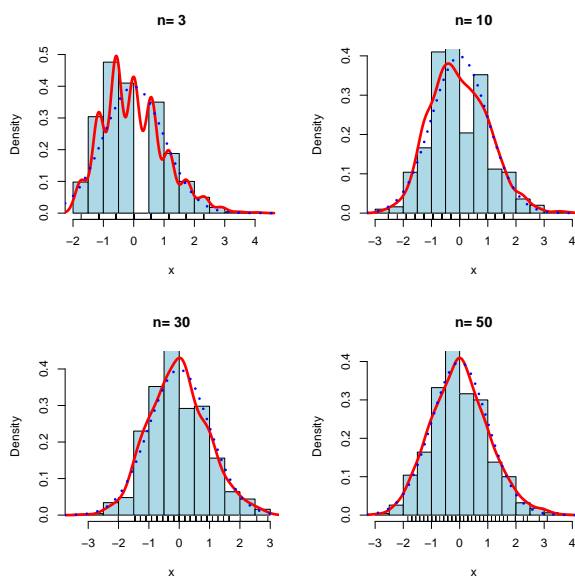


图 3.2 泊松分布的渐近正态性.

均匀分布: `unif(0,1)`

```
op <- par(mfrow=c(2,2))
limite.central( )
par(op)
```

得到图3.3.

指数分布: `exp(1)`

```
op <- par(mfrow=c(2,2))
limite.central(rexp, distpar=1, m=1, s=1)
par(op)
```

得到图3.4.

正态混合分布: $\frac{1}{2}\text{norm}(-3,1) + \frac{1}{2}\text{norm}(3,1)$

```
op <- par(mfrow=c(2,2))
```

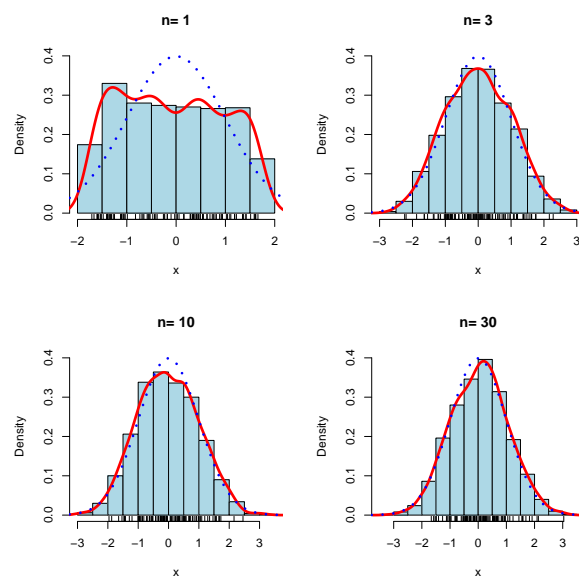


图 3.3 均匀分布的渐近正态性.

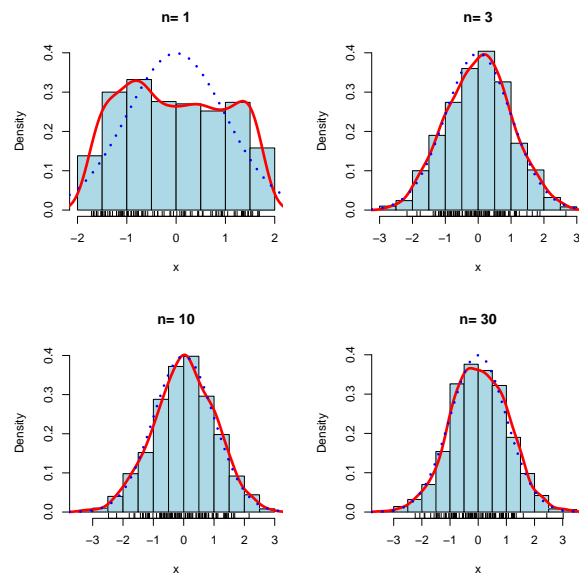


图 3.4 指数分布的渐近正态性.

```

mixn <- function (n, a=-1, b=1)
  {rnorm(n, sample(c(a,b),n,replace=T))}
limite.central(r=mixn, distpar=c(-3,3),
  m=0, s=sqrt(10), n=c(1,2,3,10))
par(op)

```

得到图3.5.

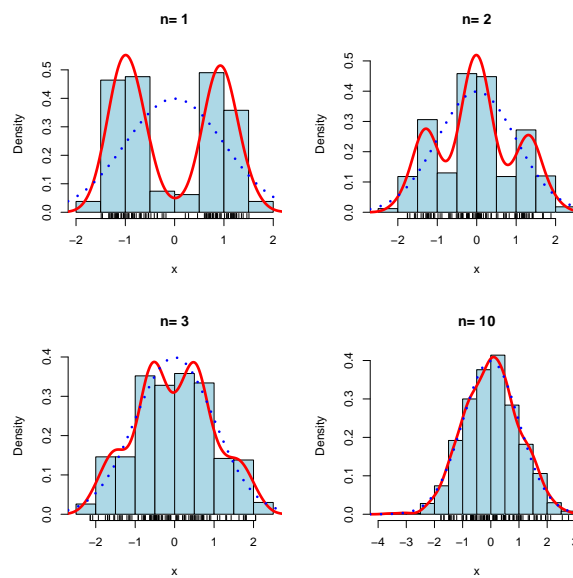


图 3.5 混合正态分布的渐近正态性.

第三章习题

3.1 从1到100个自然数中随机不放回地抽取5个数, 并求它们的和.

3.2 从一副扑克牌(52张)中随机抽5张, 求下列概率

- 抽到的是10、J、Q、K、A;
- 抽到的是同花顺.

3.3 从正态分布 $N(100, 100)$ 中随机产生1000个随机数,

- 作出这1000个正态随机数的直方图;
- 从这1000个随机数中随机有放回地抽取500个, 作出其直方图;
- 比较它们的样本均值与样本方差.

3.4 模拟随机游动: 从标准正态分布中产生1000个随机数, 并用函数`cumsum()`作出累积和, 最后使用命令`plot()`作出随机游动的示意图.

3.5 从标准正态分布中随机产生100个随机数, 由此数据求总体均值的95%置信区间, 并与理论值进行比较.

3.6 用本章给出的函数`limite.central()`, 从图形上验证当样本容量足够大时, 从贝塔分布 $\text{Beta}(1/2, 1/2)$ 抽取的样本的样本均值近似服从正态分布.

3.7 除本章给出的标准分布外, 非标准的随机变量 X 的抽样可通过格式点离散化方法实现. 设 $p(x)$ 为 X 的密度函数, 其抽样步骤如下

- 1) 在 X 的取值范围内等间隔地选取 N 个点 x_1, x_2, \dots, x_N , 例如取 $N = 1000$;
- 2) 计算 $p(x_i), i = 1, 2, \dots, N$;
- 3) 正则化 $p(x_i), i = 1, 2, \dots, N$, 使其成为离散的分布律, 即每一项除以 $\sum_{i=1}^N p(x_i)$;
- 4) 按离散分布抽样方法使用命令`sample()`从 $x_i, i = 1, 2, \dots, N$ 有放回地抽取 n 个数, 例如 $n = 1000$.

试以标准正态分布为例来说明. 为与R中的正态抽样函数`rnorm()`进行比较, 将作图区域分为左右两部分,

- 使用`rnorm()`抽取 $n = 1000$ 个标准正态随机数, 并在左侧区域画出相应的直方图和核密度估计曲线;
- 用格子点离散化抽样方法完成抽样, 并在右侧区域画出相应的直方图和核密度估计曲线, 离散化所用的 $N = 1000$, $n = 1000$, 取点范围为 $[-4, 4]$.

第四章 探索性数据分析

本章概要

- ◇ 探索性数据分析的思想
- ◇ 分布的图形概括
- ◇ 单组数据的描述性统计分析
- ◇ 多组数据的描述性统计分析
- ◇ 分组数据的描述性统计分析
- ◇ 分类数据的描述性统计分析

数据的统计分析分为描述性统计分析和统计推断两部分,前者又称为探索性统计分析,它是通过绘制统计图形、编制统计表格、计算统计量等方法来探索数据的主要分布特征,揭示其中存在的规律.探索性数据分析是进行后期统计推断的基础.本章针对不同类型的数据通过**R**介绍探索性数据分析技巧,分别从图形和描述性统计量(包括样本的均值、标准差、分位数、偏度、峰度等统计量)刻画样本的特征.

§4.1 常用分布的概率函数图

了解总体分布的形态,有助于把握样本的基本特征.我们先通过具体的例子考查第三章中提到的一些常用分布的概率函数(对于离散分布指分布律,对于连续分布指其密度函数)的图形.

二项分布

```
> n<-20
```

```
> p<-0.2
> k<-seq(0,n)
> plot(k,dbinom(k,n,p),type='h',
      main='Binomial distribution, n=20, p=0.2',xlab='k')
```

得到图4.1.

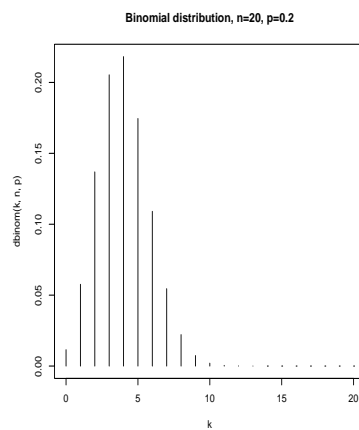


图 4.1 二项分布的分布律图

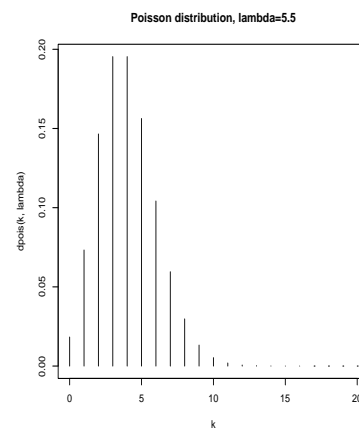


图 4.2 泊松分布的分布律图

泊松分布

```
> lambda<-4.0
> k<-seq(0,20)
> plot(k,dpois(k,lambda),type='h',
      main='Poisson distribution, lambda=5.5',xlab='k')
```

得到图4.2.

几何分布

```
> p<-0.5
> k<-seq(0,10)
> plot(k,dgeom(k,p),type='h',
      main='Geometric distribution, p=0.5',xlab='k')
```

得到图4.3.

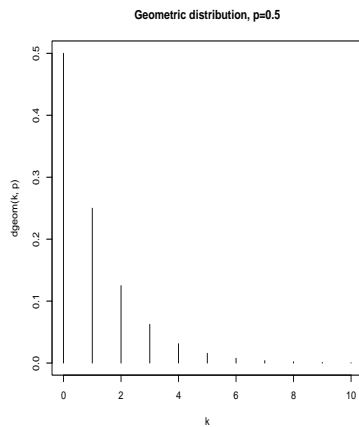


图 4.3 几何分布的分布律图

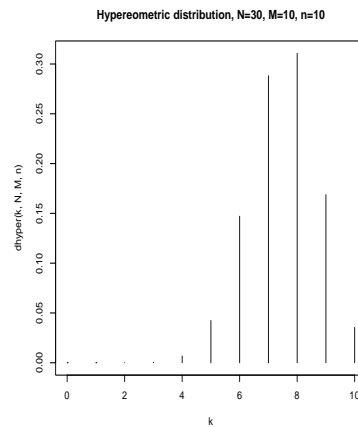


图 4.4 超几何分布分布的分布律图

超几何分布

```
> N<-30
> M<-10
> n<-10
> k<-seq(0,10)
> plot(k,dhyper(k,N,M,n),type='h',
      main='Hypergeometric distribution,
          N=30, M=10, n=10',xlab='k')
```

得到图4.4.

负二项分布

```
> n<-10
> p<-0.5
> k<-seq(0,40)
> plot(k, dnbinom(k,n,p), type='h',
      main='Negative Binomial distribution,
          n=10, p=0.5',xlab='k')
```

得到图4.5.

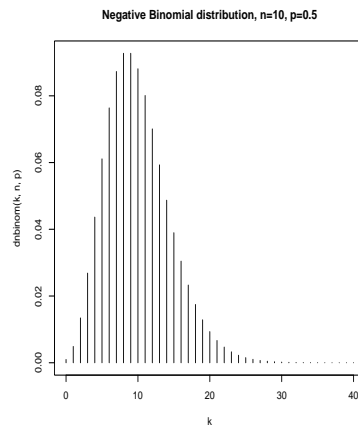


图 4.5 负二项分布的分布律图

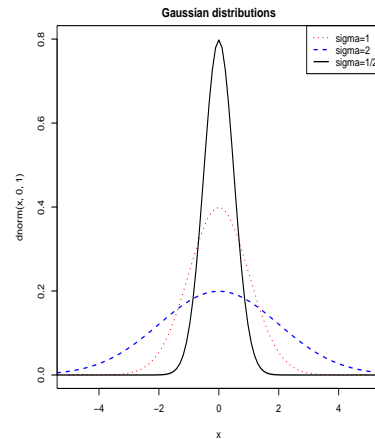


图 4.6 正态分布的密度函数图

正态分布

```
> curve(dnorm(x,0,1), xlim=c(-5,5), ylim=c(0,.8),
        col='red', lwd=2, lty=3)
> curve(dnorm(x,0,2), add=T, col='blue', lwd=2, lty=2)
> curve(dnorm(x,0,1/2), add=T, lwd=2, lty=1)
> title(main="Gaussian distributions")
> legend(par('usr')[2], par('usr')[4], xjust=1,
        c('sigma=1', 'sigma=2', 'sigma=1/2'),
        lwd=c(2,2,2),
        lty=c(3,2,1),
        col=c('red', 'blue', par("fg")))
```

得到图4.6.

t分布

```
> curve(dt(x,1), xlim=c(-3,3), ylim=c(0,.4),
        col='red', lwd=2, lty=1)
> curve(dt(x,2), add=T, col='green', lwd=2, lty=2)
```

```

> curve(dt(x,10), add=T, col='orange', lwd=2, lty=3)
> curve(dnorm(x), add=T, lwd=3, lty=4)
> title(main="Student T distributions")
> legend(par('usr')[2], par('usr')[4], xjust=1,
        c('df=1', 'df=2', 'df=10', 'Gaussian distribution'),
        lwd=c(2,2,2,2),
        lty=c(1,2,3,4),
        col=c('red', 'blue', 'green', par("fg"))))

```

得到图4.7.

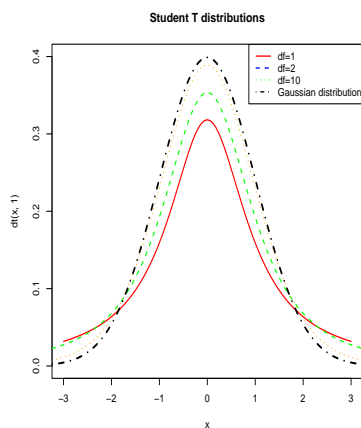


图 4.7 t分布的密度函数图

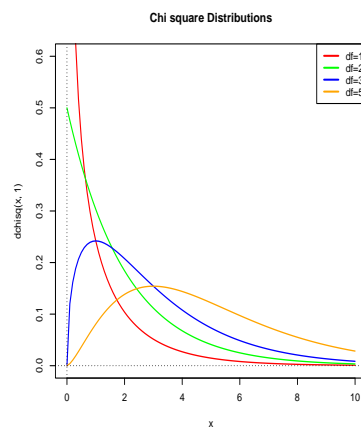


图 4.8 χ^2 分布的密度函数图

χ^2 分布

```

> curve(dchisq(x,1), xlim=c(0,10), ylim=c(0,.6), col='red', lwd=2)
> curve(dchisq(x,2), add=T, col='green', lwd=2)
> curve(dchisq(x,3), add=T, col='blue', lwd=2)
> curve(dchisq(x,5), add=T, col='orange', lwd=2)
> abline(h=0,lty=3)
> abline(v=0,lty=3)
> title(main='Chi square Distributions')
> legend(par('usr')[2], par('usr')[4], xjust=1,
        c('df=1', 'df=2', 'df=3', 'df=5'),

```

```

lwd=3, lty=1,
col=c('red', 'green', 'blue', 'orange')
)

```

得到图4.8.

F分布

```

> curve(df(x,1,1), xlim=c(0,2), ylim=c(0,.8), lty=1)
> curve(df(x,3,1), add=T, lwd=2,lty=2)
> curve(df(x,6,1), add=T, lwd=2, lty=3)
> curve(df(x,3,3), add=T, col='red', lwd=3,lty=4)
> curve(df(x,3,6), add=T, col='blue', lwd=3,lty=5)
> title(main="Fisher's F")
> legend(par('usr')[2], par('usr')[4], xjust=1,
        c('df=(1,1)', 'df=(3,1)', 'df=(6,1)',
          'df=(3,3)', 'df=(3,6)'),
        lwd=c(1,2,2,3,3),
        lty=c(1,2,3,4,5),
        col=c(par("fg"), par("fg"), par("fg"), 'red', 'blue'))

```

得到图4.9.

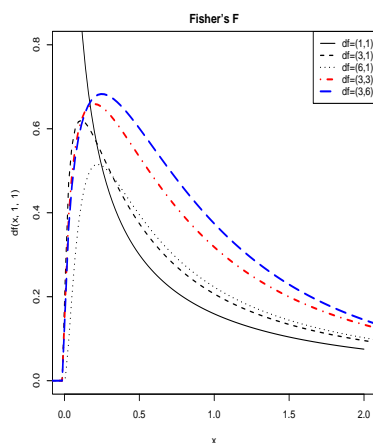


图 4.9 F 分布的密度函数图

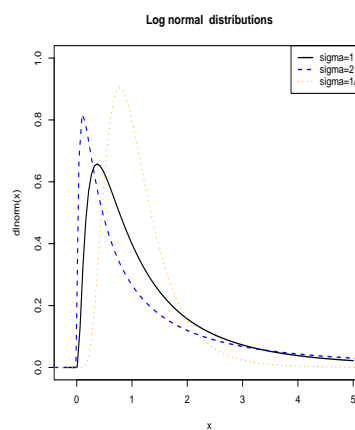


图 4.10 对数正态分布的密度函数图

对数正态分布

```
> curve(dlnorm(x), xlim=c(-.2,5), ylim=c(0,1.0), lwd=2)
> curve(dlnorm(x,0,3/2), add=T, col='blue', lwd=2, lty=2)
> curve(dlnorm(x,0,1/2), add=T, col='orange', lwd=2, lty=3)
> title(main="Log normal distributions")
> legend(par('usr')[2], par('usr')[4], xjust=1,
        c('sigma=1', 'sigma=2', 'sigma=1/2'),
        lwd=c(2,2,2),
        lty=c(1,2,3),
        col=c(par("fg"), 'blue', 'orange' ))
```

得到图4.10.

柯西分布

```
> curve(dcauchy(x),xlim=c(-5,5), ylim=c(0,.5), lwd=3)
> curve(dnorm(x), add=T, col='red', lty=2)
> legend(par('usr')[2], par('usr')[4], xjust=1,
        c('Cauchy distribution', 'Gaussian distribution'),
        lwd=c(3,1),
        lty=c(1,2),
        col=c(par("fg"), 'red'))
```

得到图4.11.

威布尔分布

```
> curve(dexp(x), xlim=c(0,3), ylim=c(0,2))
> curve(dweibull(x,1), lty=3, lwd=3, add=T)
> curve(dweibull(x,2), col='red', add=T)
> curve(dweibull(x,.8), col='blue', add=T)
> title(main="Weibull Probability Distribution Function")
> legend(par('usr')[2], par('usr')[4], xjust=1,
        c('Exponential', 'Weibull, shape=1',
          'Weibull, shape=2', 'Weibull, shape=.8'),
```

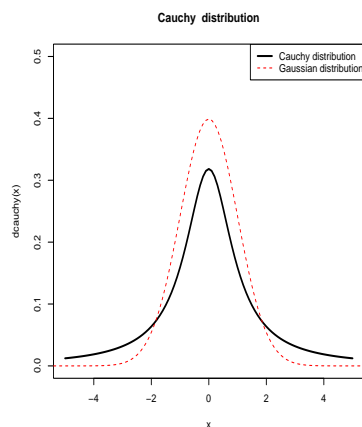


图 4.11 柯西分布的密度函数图

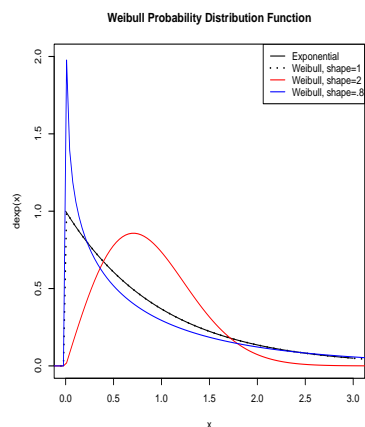


图 4.12 威布尔分布的密度函数图

```
lwd=c(1,3,1,1),
lty=c(1,3,1,1),
col=c(par("fg"), par("fg"), 'red', 'blue'))
```

得到图4.12.

伽码分布

```
> curve( dgamma(x,1,1), xlim=c(0,5), lwd=2, lty=1 )
> curve( dgamma(x,2,1), add=T, col='red', lwd=2, lty=2 )
> curve( dgamma(x,3,1), add=T, col='green', lwd=2, lty=3 )
> curve( dgamma(x,4,1), add=T, col='blue', lwd=2, lty=4 )
> curve( dgamma(x,5,1), add=T, col='orange', lwd=2, lty=5 )
> title(main="Gamma distributions")
> legend(par('usr')[2], par('usr')[4], xjust=1,
        c('k=1 (Exponential distribution)',
          'k=2', 'k=3', 'k=4', 'k=5'),
        lwd=c(2,2,2,2,2),
        lty=c(1,2,3,4,5),
        col=c(par('fg'), 'red', 'green', 'blue', 'orange')) )
```

得到图4.13.

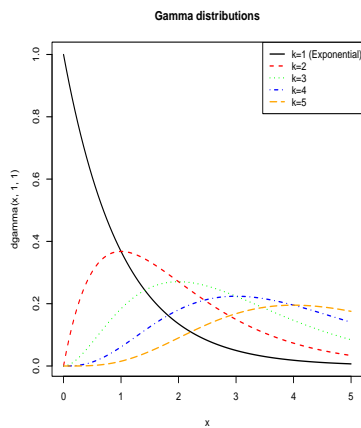


图 4.13 伽玛分布的密度函数图

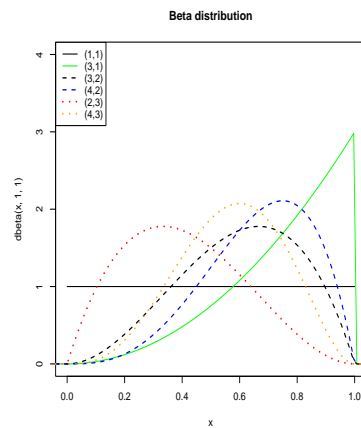


图 4.14 贝塔分布的密度函数图

贝塔分布

```
> curve( dbeta(x,1,1), xlim=c(0,1), ylim=c(0,4) )
> curve( dbeta(x,3,1), add=T, col='green' )
> curve( dbeta(x,3,2), add=T, lty=2, lwd=2 )
> curve( dbeta(x,4,2), add=T, lty=2, lwd=2, col='blue' )
> curve( dbeta(x,2,3), add=T, lty=3, lwd=3, col='red' )
> curve( dbeta(x,4,3), add=T, lty=3, lwd=3, col='orange' )
> title(main="Beta distributions")
> legend(par('usr')[1], par('usr')[4], xjust=0,
        c('(1,1)', '(3,1)', '(3,2)',
          '(4,2)', '(2,3)', '(4,3)' ),
        lwd=c(1,1, 2,2, 3,3),
        lty=c(1,1, 2,2, 3,3),
        col=c(par('fg'), 'green', par('fg'),
              'blue', 'red', 'orange' ))
```

得到图4.14.

§4.2 直方图与密度函数的估计

4.2.1 直方图

直方图是探索性数据分析的基本工具,它给出了数据的频率分布图形,在组距相等场合下常用宽度相等的长条矩形表示,矩形的高低表示频率的大小.在图形上,横坐标表示所关心变量的取值区间,纵坐标表示频率(或频数)的大小,这样就得到频数(或频数)直方图.图形的形状与我们选择的各组区间端点有关,故选择区间端点时我们要谨慎.

R中使用函数`hist()`来画直方图,其常用的调用格式如下:

—— `hist()` 的调用格式 ——

```
hist(x, breaks = "Sturges", freq = NULL, probability = !freq,
     col = NULL,
     main = paste("Histogram of" , xname),
     xlim = range(breaks), ylim = NULL,
     xlab = xname, ylab,
     axes = TRUE, nclass = NULL)
```

说明:若选项`breaks`取向量,则用于指明直方图区间的分割位置;若取正整数,则用于指定直方图的小区间数. `freq`取T表示使用频数画直方图,取F则使用频率画直方图. `probability`与`freq`恰好相反. `col`用于指明小矩形的颜色. 其它选项可参考`hist()`的帮助说明. 后面我们还将给出`hist()`的二种拓展.

4.2.2 核密度估计

样本的直方图粗略地描述了样本的分布,我们还可以用函数`density()`得到样本的核密度估计值,并用`lines()`得到密度估计的曲线. `density()`常用的调用格式如下:

—— `density()` 的调用格式 ——

```
density(x, bw = "nrd0",
        kernel = c("gaussian", "epanechnikov", "rectangular",
                   "triangular", "biweight", "cosine", "optcosine"),
        n = 512, from, to)
```

说明:选项`bw`指定核密度估计的窗宽,也用字符串表示窗宽选择规则,具体可参考函数`bw.nrd()`. `kernel`为核密度估计所使用的光滑化函数,缺省为正态

核函数. `n`给出等间隔的核密度估计点. `from`与`to`分别给出需要计算核密度估计的左右端点. 其它选项可参考`density()`的在线帮助¹.

下面看两个模拟例子.

例 4.2.1 从二项分布`binom(100,0.9)`中抽取容量为 $N=100000$ 的样本. 试作出它的直方图及核密度估计曲线.

```
> N <- 100000
> n <- 100
> p <- .9
> x <- rbinom(N,n,p)
> hist(x,
      xlim=c(min(x),max(x)), probability=T,
      nclass=max(x)-min(x)+1, col='lightblue',
      main='Binomial distribution, n=100, p=.5')
> lines(density(x,bw=1), col='red', lwd=3)
```

得到图4.16.

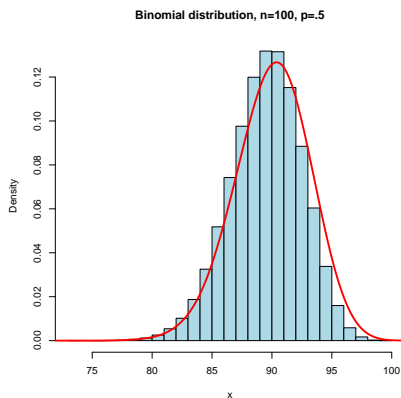


图 4.15 二项分布的样本的直方图与核密度函数图

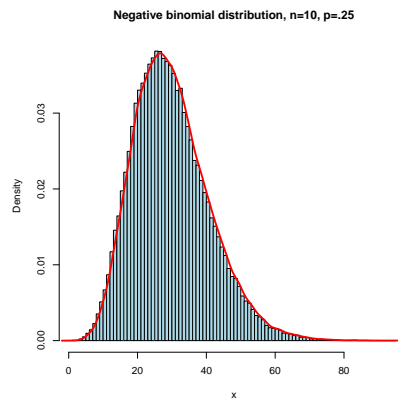


图 4.16 负二项分布的样本的直方图与核密度函数图

例 4.2.2 从负二项分布`nbinom(10,0.25)`中抽取容量为 $N=100000$ 的样本. 试作出它的直方图及核密度估计曲线.

¹样本的密度函数估计也可使用局部多项式估计程序包`locfit`中的`density.lf()`函数实现.

```
> N <- 100000
> x <- rnbinoM(N, 10, .25)
> hist(x,
      xlim=c(min(x),max(x)), probability=T,
      nclass=max(x)-min(x)+1, col='lightblue',
      main='Negative binomial distribution, n=10, p=.25')
lines(density(x,bw=1), col='red', lwd=3)
```

得到图4.17.

§4.3 单组数据的描述性统计分析

4.3.1 单组数据的图形描述

单组数据的分布可以通过上面介绍的直方图以及茎叶图和框须图考查.

例 4.3.1 程序包DAAG中有内嵌数据集“possum”，它包括了从维多利亚南部到皇后区的七个地区的104只负鼠(possum)的年龄、尾巴的长度、总长度等9个特征值，我们仅考虑43只雌性负鼠的特征值，我们建立子集fpossum，考查雌性负鼠(fpossum)的总长度的频率分布.

直方图

```
> library(DAAG)
> data(possum)
> fpossum <- possum[possum$sex=="f",]
> par(mfrow=c(1,2))
> attach(fpossum)
> hist(totlngth,breaks=72.5+(0:5)*5,
      ylim=c(0,22), xlab="total length",
      main="A:Breaks at 72.5,77.5...")
> hist(totlngth,breaks=75+(0:5)*5,
      ylim=c(0,22), xlab="total length",
      main="B:Breaks at 75,80...")
```

得到图4.15. 两个图的唯一不同之处是选择的区间端点不同,我们可以看到左

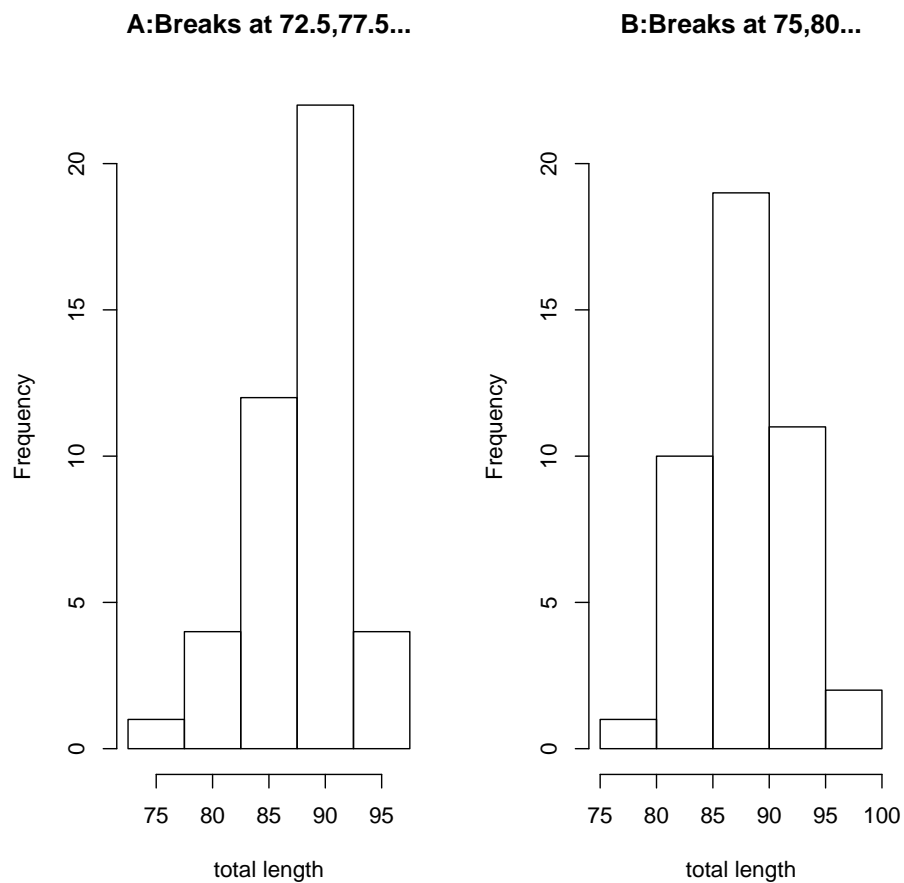


图 4.17 雌性负鼠的直方图.

左边的图不对称,而右边显示该分布是对称的.

茎叶图

茎叶图也是考查数据分布的重要方法,我们仍然考虑上面的雌性负鼠的总长度.

```
> stem(fpossum$totlngth)
```

得到

```

The decimal point is at the |
74 | 0
76 |
78 |
80 | 05
82 | 0500
84 | 05005
86 | 05505
88 | 0005500005555
90 | 5550055
92 | 000
94 | 05
96 | 5

```

说明: 左边茎是长度(厘米)的整数部分, 右边是小数点后边的部分, 由于数据采用了近似, 所以右边只有0与5, 显然叶的部分是左边长度(厘米)整数部分的频数. 图中有43个的数据, 中位数是第22个. 可知从上至下第22个叶对应的茎是88, 叶是5, 因此样本中位数应该是88.5. 茎叶图的外观很像横放的直方图, 但茎叶图中的叶增加了具体的数值, 从而保留了数据更多的信息.

框须图

框须图, 或称为盒形图, 是五数(最小值、第三四分位数、中位数、第一四分位数、最大值)的图形概括, 也是考查数据分析的一种有效的工具, 它可用来对数据分布的形状进行大致的判断. 在R中使用函数`boxplot()`作盒形图. `boxplot()`的调用格式如下:

`boxplot()`的调用格式

```
boxplot(formula, data = NULL, ..., subset, na.action = NULL)
```

说明: `formula` 是指明盒形图的作图规则($y \sim \text{grp}$, 表示数值变量 y 根据因子 grp 分类), `data`说明数据的来源.

```

> library(DAAG)
> data(possum)
> fpossum <- possum[possum$sex=="f",]
> boxplot(fpossum$totlngth)

```


得到图4.18. 箱子中的五根横线对应的坐标分别是最小值,第一4分位数, 中位数, 第三4分位数和最大值.

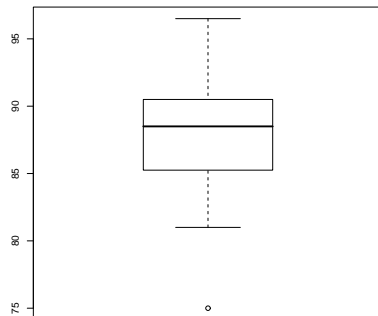


图 4.18 雌性负鼠的框须图

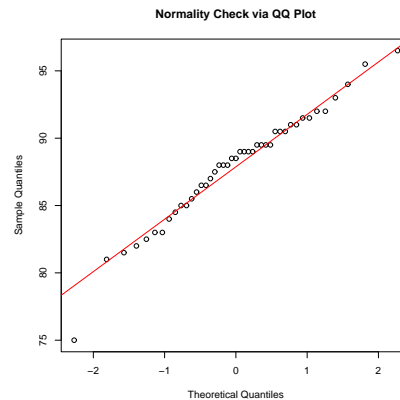


图 4.19 雌性负鼠的QQ图

正态性检验

1) 使用QQ图

```
> qqnorm(fpossum$totlngth,
          main="Normality Check via QQ Plot")
> qqline(fpossum$totlngth, col='red')
```

得到图4.19. 图4.19表明数据与正态性略有差异, 特别在图形的中部.

2) 与正态密度函数比较

```
> dens <- density(totlngth)
> xlim <- range(dens$x); ylim<-range(dens$y)
> par(mfrow=c(1,2))
> hist(totlngth,breaks=72.5+(0:5)*5,
       xlim=xlim,ylim=ylim,
       probability=T, xlab="total length",
       main="A:Breaks at 72.5,77.5...")
> lines(dens,col=par('fg'),lty=2)
> m <- mean(totlngth)
```

```

> s <- sd(totlngth)
> curve( dnorm(x, m, s), col='red', add=T)
> hist(totlngth,breaks=75+(0:5)*5,
      xlim=xlim,ylim=ylim,
      probability=T, xlab="total length",
      main="B:Breaks at 75,80...")
> lines(dens,col=par('fg'),lty=2)
> m <- mean(totlngth)
> s <- sd(totlngth)
> curve( dnorm(x, m, s), col='red', add=T)

```

得到图4.20. 图4.20表明数据totlngth与正态性也略有差异. 进一步需要使用统计量进行正态性检验.

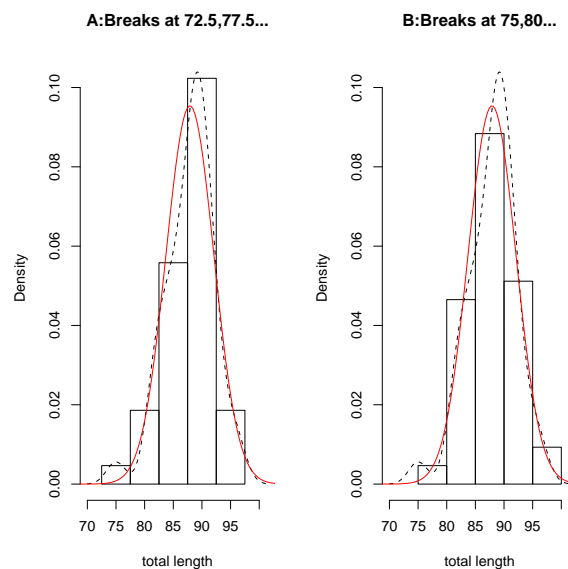


图 4.20 雌性负鼠的核密度与正态分布的比较.

3) 使用经验分布函数

```

> x <- sort(totlngth)
> n <- length(x)
> y <- (1:n)/n

```

```
> m <- mean(totlngth)
> s <- sd(totlngth)
> plot(x,y, type='s', main="empirical cdf of ")
> curve(pnorm(x,m,s),col='red', lwd=2, add=T)
```

得到图4.21. 结论与前面类似.

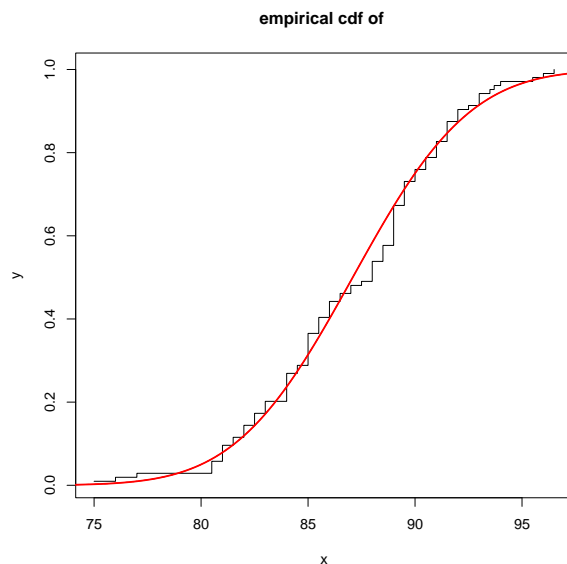


图 4.21 雌性负鼠的经验分布.

4.3.2 单组数据的描述性统计

样本来自总体, 样本的观测值中含有总体各方面的信息, 但这些信息较为分散, 有时显得杂乱无章. 为将这些分散在样本中的有关总体的信息集中起来以反映总体的各种特征, 需要对样本进行加工得到统计量. 均值、标准差、五数(最小值、第三4分位数、中位数、第一4分位数、最大值)是数据的主要的统计量, 他们对数据的进一步分析很有帮助.

总体描述

在R中, 函数summary()可以计算出单组数据的均值和五数. 仍然用上一节的例子, 考虑雌性负鼠的总长度.

```
> summary(fpossum$totlngth)
```

得到

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
75.00	85.25	88.50	87.91	90.50	96.50

如果只需要均值可以利用函数`mean()`实现

```
> mean(fpossum$totlngth)
[1] 87.90698
```

五数及样本分位数概括

计算五数用函数`fivenum()`。若要得到分位数用函数`quantile()`，计算中位数使用函数`median()`，最大值使用函数`max()`，最小值使用函数`min()`。我们在第二章中提过，计算更多概率值的样本分位数，可使用选项`probs`。以雌性负鼠的总长度为例：

```
> fivenum(fpossum$totlngth)
[1] 75.00 85.25 88.50 90.50 96.50
> quantile(fpossum$totlngth)
 0%  25%  50%  75% 100%
75.00 85.25 88.50 90.50 96.50
> quantile(fpossum$totlngth ,prob=c(0.25,0.5,0.75))
 25%  50%  75%
85.25 88.50 90.50
> median(fpossum$totlngth)
[1] 88.5
> max(fpossum$totlngth)
[1] 96.5
> min(fpossum$totlngth)
[1] 75
```

离差的概括

样本的平均水平可以用上面介绍的平均值函数`mean()`和中位数函数`median()`来计算。样本的变异程度可以用极值(`max()`-`min()`)、四分位极

值函数(`IQR()`)、标准差函数(`sd()`)、方差函数(`var()`)和绝对离差函数(`mad()`)来表示. 方差函数`var()`也可用于计算两个向量协方差或一个矩阵的协方差阵. 对于 $x = (x_1, \dots, x_n)$, `sd()`的定义为

$$sd(x) = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}.$$

`mad()`在R中的定义为

$$1.4826 * \text{median}(\text{abs}(x - \text{median}(x)))$$

其中系数1.4826约等于 $1/\text{qnorm}(3/4)$, 目的是为了使得`mad(x)`作为方差的估计具有一致性(在正态或大样本下). 仍以雌性负鼠的总长度为例:

```
> max(fpossum$totlngth)-min(fpossum$totlngth)
[1] 21.5
> IQR(fpossum$totlngth)
[1] 5.25
> sd(fpossum$totlngth)
[1] 4.182241
> sd(fpossum$totlngth)^2
[1] 17.49114
> var(fpossum$totlngth)
var(fpossum$totlngth)
> mad(fpossum$totlngth)
[1] 3.7065
```

样本偏度系数和峰度系数

设随机变量 X 的三阶矩存在, 则称比值

$$\beta_1 = \frac{E(X - E(X))^3}{[E(X - E(X))^2]^{3/2}} = \frac{\nu_3}{(\nu_2)^{3/2}}$$

为 X 的偏度系数. $\beta_1 > 0$ 时分布为正偏(或右偏); $\beta_1 = 0$ 时分布关于均值对称; $\beta_1 < 0$ 时分布为负偏(或左偏). 用样本的中心矩代替总体的中心矩就可得到样本的偏度系数.

设随机变量 X 的四阶矩存在,则称比值

$$\beta_2 = \frac{E(X - E(X))^4}{[E(X - E(X))^2]^2} - 3 = \frac{\nu_4}{(\nu_2)^2} - 3$$

为 X 的峰度系数. 峰度系数刻画的是分布的峰度, $\beta_2 > 0$ 时标准化后的分布形状比高斯分布更尖锐,称为高峰度; $\beta_2 = 0$ 时标准化后的分布形状与高斯分布相当; $\beta_2 < 0$ 时标准化后的分布形状比高斯分布更平坦,称为低峰度. 用样本的中心矩代替总体的中心矩就可得到样本的偏度系数.

R的扩展统计程序包**fBasics**提供了函数**skewness()**用来求样本的偏度, 函数**kurtosis()**用来求样本的峰度. 对于雌性负鼠的总长度有

```
> library(fBasics)
> skewness(fpossum$totlngth)
[1] -0.54838
> kurtosis(fpossum$totlngth)
[1] 0.6170082
```

另外, **fBasics**程序包的函数**basicStats()**提供了几乎上面所有的统计特征量.

§4.4 多组数据的描述性统计分析

4.4.1 两组数据的图形概括

散点图

在两组数据的图形展示中, 散点图是简单而重要的工具, 因为它能清楚地描述两组数据的关系. 下面我们来看一个例子.

例 4.4.1 在**R**的程序包**DAAG**中有数据集**cars**, 使用下边的命令得到数据集.

```
> library(DAAG)
> data(cars)
> cars
  speed dist
```

```
1      4      2
2      4     10
3      7      4
... ..
48     24     93
49     24    120
50     25     85
```

我们希望估计速度(speed)与终止距离(dist)之间的关系, 先考查它们之间的散点图, 命令

```
> plot(cars$dist ~ cars$speed,
       xlab = "Speed (mph)",
       ylab = "Stopping distance (ft)")
> lines(lowess(cars$speed, cars$dist),lwd=2)
```

得到图4.22. 图4.22表明speed和dist基本呈现线性相依关系. 所以散点图在描

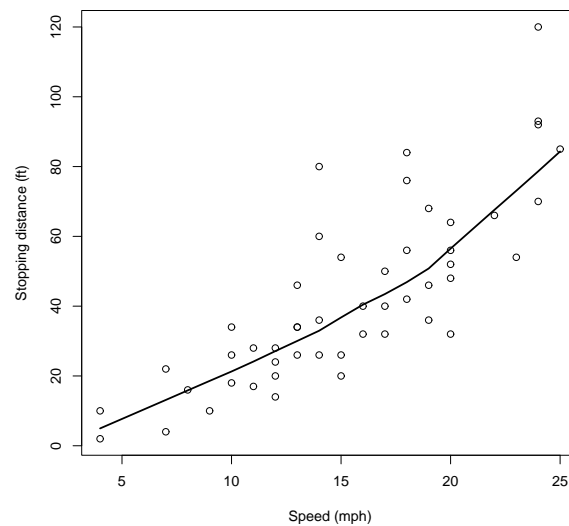


图 4.22 speed与dist的散点图.

述二维数据的关系方面很重要.

注意到我们用一条非线性的特殊曲线来拟和这种关系,调用了函数`lowess()`。在R中,有两个函数可以实现这个功能,一个是`lowess()`,另一个是`loess()`,前者只能适用于二维的情况,而`loess()`可以处理多维的情况。`lowess()`的具体的调用格式如下:

`lowess()`的调用格式

```
lowess(x, y = NULL, f = 2/3, iter = 3,
       delta = 0.01 * diff(range(xy$x[o])))
```

在散点图中加入拟和曲线对于我们认识总体的特征很有帮助。

进一步,通过函数`rug()`可以在横轴和纵轴上标明数据的具体位置。

```
> rug(side=2, jitter(cars$dist, 20))
> rug(side=1, jitter(cars$speed, 5))
```

得到图4.23。

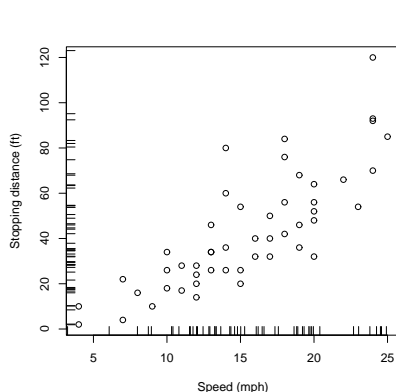


图 4.23 带rug的散点图

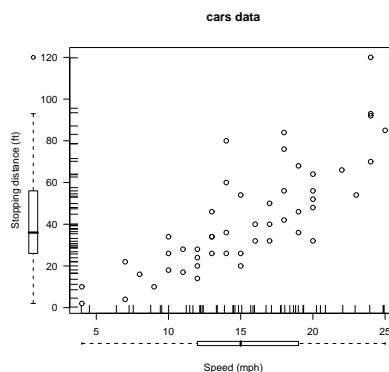


图 4.24 加上箱形图的散点图

我们也可以在数轴两边加上单变量的箱形图。

```
> op <- par( )
> layout(matrix(c(2,1,0,3), 2, 2, byrow=T ), c(1,6), c(4,1))
> par(mar=c(1,1,5,2))
> plot(cars$dist ~ cars$speed,
       xlab='', ylab='', las = 1)
```



```
> rug(side=1, jitter(cars$speed, 5) )
> rug(side=2, jitter(cars$dist, 20) )
> title(main = "cars data")
> par(mar=c(1,2,5,1))
> boxplot(cars$dist, axes=F)
> title(ylab='Stopping distance (ft)', line=0)
> par(mar=c(5,1,1,2))
> boxplot(cars$speed, horizontal=T, axes=F)
> title(xlab='Speed (mph)', line=1)
> par(op)
```

运行得到图4.24. 这样我们既可以了解两个变量的统计量也可以看出两变量之间的关系.

等高线图

有时候数据太多太集中,散点图上的信息不容易看出来. 例如由

```
> library(chplot)
> data(hdr)
> x<- hdr$age
> y<- log(hdr$income)
> plot(x,y)
```

得到的图4.25. 这时我们要借助于二维的密度估计来认识图形. 首先使用MASS程序包中的二维核密度估计函数kde2d()来估计这个二维数据的密度函数, 再利用函数contour()画出密度的等高曲线图.

```
> library(MASS)
> z <- kde2d(x,y)
> contour(z, col = "red", drawlabels = FALSE,
          main = "Density estimation: contour plot")
```

运行得到图4.26.

三维透视图

我们也可以利用函数persp()作出三维透视图, 这样看更形象.

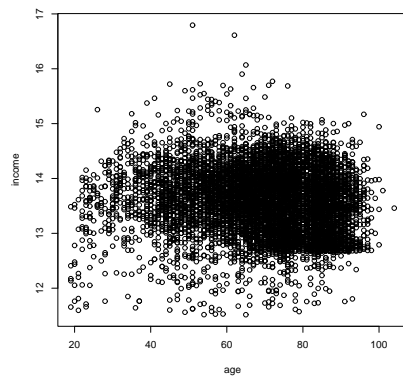


图 4.25 age与income的散点图

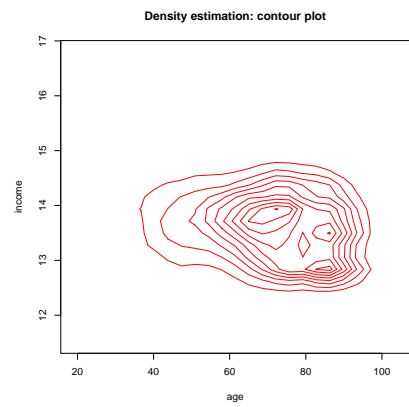


图 4.26 age与income的等高线图

```
> persp(z, main = "Density estimation: perspective plot")
```

运行得到图4.27.

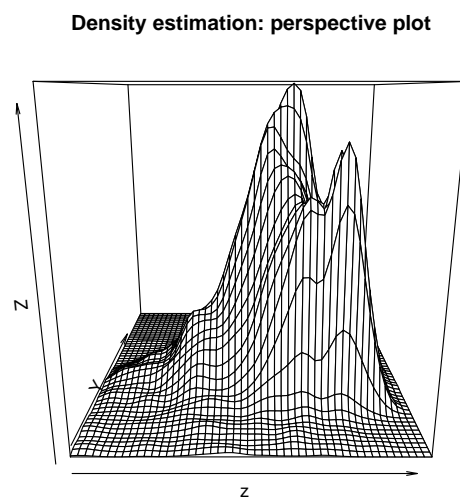


图 4.27 三维图形.

数据的变换

当直接用原数据得不到有意义的图形时, 可以对数值进行变换以得到有意义的图形, 最常用的是对数变换、倒数变换、指数变换和更为一般的Box-Cox变换:

$$f(x) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{如果 } \lambda \neq 0, \\ \log(y), & \text{如果 } \lambda = 0. \end{cases}$$

我们用程序包MASS中的数据集合Animal来举例说明.

例 4.4.2 首先调出数据集Animal

```
> library(MASS)
> data(Animals)
> Animals
```

输出数据结果如下:

	body	brain
Mountain beaver	1.350	8.1
Cow	465.000	423.0
Grey wolf	36.330	119.5
Goat	27.660	115.0
Guinea pig	1.040	5.5
Dipliodocus	11700.000	50.0
...		
Pig	192.000	180.0

我们在R中画两个图, 一个使用原始的数据, 另一个对原来的数值取对数.

```
> par(mfrow=c(1,2))
> plot(brain~body,data=Animals)
> plot(log(brain)~log(body),data=Animals)
```

得到图4.28. 可以看到图4.28左侧的散点图没有价值, 而从右侧的散点图可以看出两组数据在取对数后呈现明显的线性相依关系. 对两组数据取对数的技巧在绘图中很常见, 生活中许多数据成指数上升趋势, 比如细胞繁殖, 这种数据

取对数后就呈线性上升趋势. 因此, 对数据作对数处理(或更为一般的变换)很有意义.

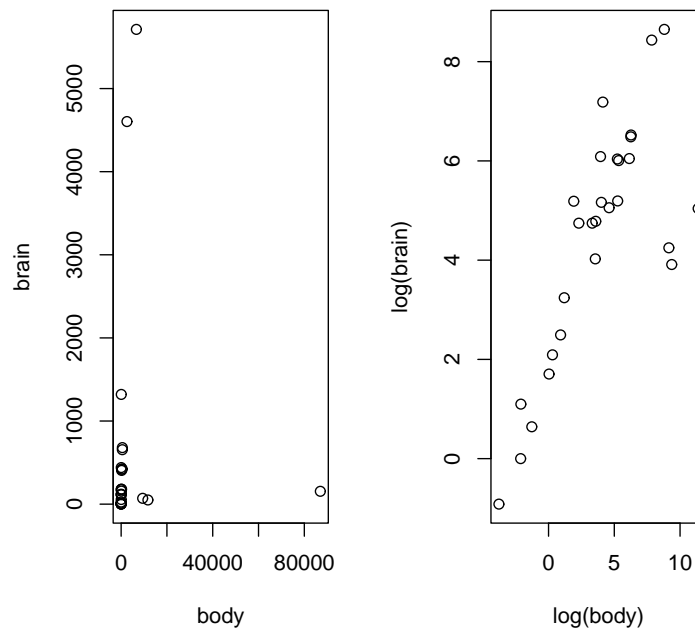


图 4.28 数据变换比较图.

4.4.2 多组数据的图形描述

对多组数据, 我们给出3种作图的方法(函数): `pairs()`或`plot()`, `matplot()`和`boxplot()`. 它们都可以看成一维或二维画图函数的延伸. 我们仅通一个例子加以说明, 具有使方法可参考相应的帮助文件.

例 4.4.3

```
> n<-10  
> d<-data.frame(y1 = abs(rnorm(n)),  
                y2 = abs(rnorm(n)),  
                y3 = abs(rnorm(n)),  
                y4 = abs(rnorm(n)),
```

```
y5 = abs(rnorm(n))  
)
```

散点图

多组数据的散点图就是不同变量的散点图像矩阵一样放在一起, 使用的函数为`pairs()`, 也可直接使用散点图函数`plot()`. 运行

```
>plot(d) # 或者 pairs(d)
```

得到图4.29.

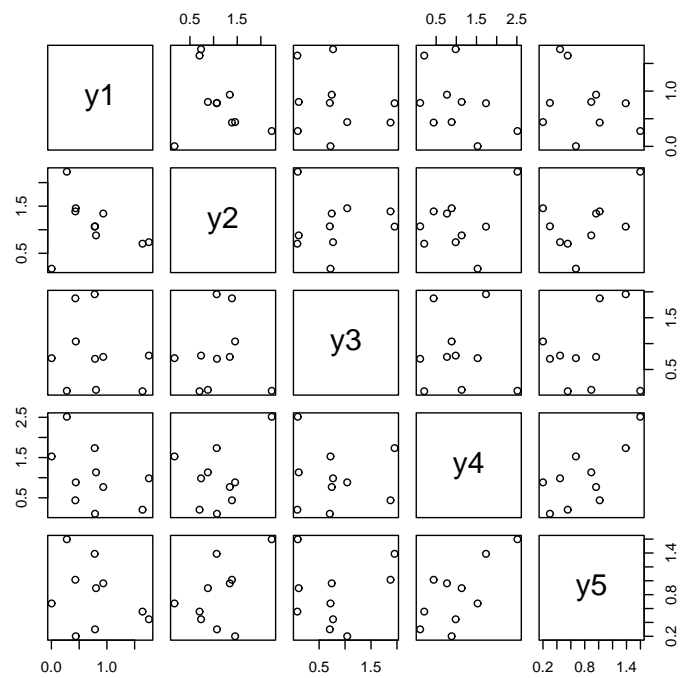


图 4.29 多组数据的散点图.

矩阵图

`matplot()` 在处理多组数据时很好用. 它与散点图矩阵的区别是将各个散点图放在同一个作图区域中. 对于上面的模拟数据运行

```
> matplot(d, type = 'l', ylab = "", main = "Matplot")
```

得到图4.30.

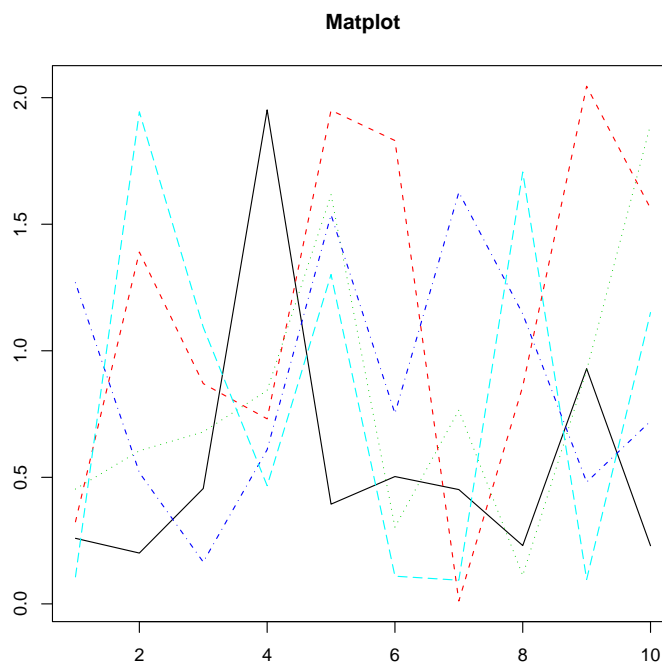


图 4.30 多组数据的matplot图.

框须图

使用函数`boxplot()`可在同一个作图区域画出各组数的框须图(盒形图), 对于上面的数据运行

```
> boxplot(d)
```

得到图4.31.

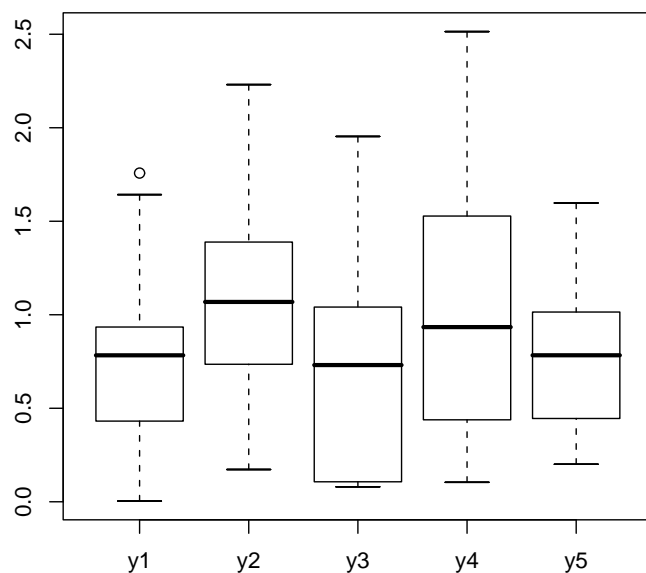


图 4.31 多组数据的boxplot图.

分组数据比较特殊,它既含有定性的变量,又含有数值型变量,而上面所说的多组数据,我们仅局限于数值型的观测.我们将在后面一节专门给出带定性变量的分组数据的描述性统计分析.

4.4.3 多组数据的描述性统计

多组数据的概述

对多组数据进行概述与单组数据情形类似,直接使用`summary()`可以得到各组数据的均值和五数.先看一个例子

例 4.4.4 程序包`datasets`中数据框`state.x77`描述了美国50个州的人口数、人均收入、人均寿命、一年中有雾的天数等情况.数据如下:

```
> state.x77

      Population Income Illiteracy Life Exp ...
Alabama      3615   3624         2.1   69.05 ...
Alaska        365   6315         1.5   69.31 ...
Arizona      2212   4530         1.8   70.55 ...
... ..
Wisconsin     4589   4468         0.7   72.48 ...
Wyoming       376   4566         0.6   70.29 ...
```

使用函数`summary()`概括`state.x77`, 结果如下:

```
> summary(state.x77)

      Population      Income      Illiteracy      Life Exp
Min.   : 365   Min.   :3098   Min.   :0.500   Min.   :67.96
1st Qu.:1080   1st Qu.:3993   1st Qu.:0.625   1st Qu.:70.12
Median :2839   Median :4519   Median :0.950   Median :70.67
Mean   :4246   Mean   :4436   Mean   :1.170   Mean   :70.88
3rd Qu.:4969   3rd Qu.:4814   3rd Qu.:1.575   3rd Qu.:71.89
Max.   :21198  Max.   :6315   Max.   :2.800   Max.   :73.60

      Murder      HS Grad      Frost      Area
Min.   : 1.400   Min.   :37.80   Min.   : 0.00   Min.   : 1049
1st Qu.: 4.350   1st Qu.:48.05   1st Qu.: 66.25   1st Qu.: 36985
Median : 6.850   Median :53.25   Median :114.50   Median : 54277
Mean   : 7.378   Mean   :53.11   Mean   :104.46   Mean   : 70736
3rd Qu.:10.675   3rd Qu.:59.15   3rd Qu.:139.75   3rd Qu.: 81163
Max.   :15.100   Max.   :67.30   Max.   :188.00   Max.   :566432
```

为了统计不同地区(Northeast, South, North Central, West)的这几个变量的均值(或中位数、分位数) 可以使用分组概括函数`aggregate()`, 其调用格式如下:

`aggregate()` 的调用格式

```
aggregate(x, by, FUN, ...)
```

说明: `x`是数据框, `by`指定分组变量, `fun`是用于计算的统计函数. 如果计算均值, `fun`为`mean`. 接着上面的例子计算各个地区各个变量的均值:


```
> aggregate(state.x77, list(Region = state.region), mean)
      Region      Population      Income      Illiteracy      LifeExp ...
1 Northeast      5495.111      4570.222      1.000000      71.26444 ...
2 South          4208.125      4011.938      1.737500      69.70625 ...
3 North Central  4803.000      4611.083      0.700000      71.76667 ...
4 West           2915.308      4702.615      1.023077      71.23462 ...
```

同样, 根据不同地区和是否一年中有雾的天数超过130来统计这几个变量的均值:

```
aggregate(state.x77, list(Region = state.region,
                          Cold = state.x77[, "Frost"] > 130), mean)
      Region Cold Population      Income      Illiteracy      Life Exp ...
1 Northeast FALSE  8802.8000  4780.400      1.1800000  71.12800 ...
2 South      FALSE  4208.1250  4011.938      1.7375000  69.70625 ...
3 North Central FALSE  7233.8333  4633.333      0.7833333  70.95667 ...
4 West       FALSE  4582.5714  4550.143      1.2571429  71.70000 ...
5 Northeast  TRUE   1360.5000  4307.500      0.7750000  71.43500 ...
6 North Central TRUE   2372.1667  4588.833      0.6166667  72.57667 ...
7 West       TRUE    970.1667  4880.500      0.7500000  70.69167 ...
```

注: Cold为TRUE表示该地区一年有雾的天数超过130天; Cold为FALSE 表示该地区一年有雾的天数没有超过130天。

标准差与协方差阵的计算

变量标准差的计算仍然使用函数sd()

```
> options(digits=3)
> sd(state.x77)
Population Income Illiteracy Life Exp Murder HS Grad Frost      Area
4464.49 614.47      0.61      1.34      3.69      8.08 51.98 85327.30
```

函数var()应用在多组数据中计算的是协方差阵:

```
> var(state.x77)
      Population      Income      Illiteracy      Life Exp ...
```

```
Population 19931684 571230 292.868 -4.08e+02 ...
Income      571230 377573 -163.702 2.81e+02 ...
Illiteracy   293 -164 0.372 -4.82e-01 ...
Life Exp    -408 281 -0.482 1.80e+00 ...
Murder       5664 -522 1.582 -3.87e+00 ...
HS Grad     -3552 3077 -3.235 6.31e+00 ...
Frost       -77082 7228 -21.290 1.83e+01 ...
Area        8587917 19049014 4018.337 -1.23e+04 ...
```

同上,我们可以用函数`aggregate()`分别计算不同区域的标准差:

```
> aggregate(state.x77, list(Region = state.region), sd)
      Region Population Income Illiteracy Life Exp ...
1 Northeast      6080    559    0.278    0.744 ...
2 South         2780    605    0.552    1.022 ...
3 North Central  3703    283    0.141    1.037 ...
4 West          5579    664    0.608    1.352 ...
```

相关系数的计算

散点图让我们对两组数据的线性相依关系有了直观的认识, 皮尔逊(Pearson)相关系数可以度量这种线性相关性程度. 如果数据呈现的不是线性关系, 而是单调的, 这时可使用斯皮尔曼(Spearman) 或者肯德尔(Kendall)相关系数, 因为它们描述的是秩相关性. 在R中我们使用函数`cor()`计算相关系数或相关系数矩阵, 其调用格式如下:

cor() 的调用格式

```
cor(x, y = NULL, use = "all.obs", method =
    c("pearson", "kendall", "spearman"))
```

例如, 我们计算下面二个向量 x 与 y 之间的三个相关系数:

```
> x<-c(44.4, 45.9, 46.0, 46.5, 46.7, 47, 48.7, 49.2, 60.1)
> y<-c(2.6, 10.1, 11.5, 30.0, 32.6, 50.0, 55.2, 85.8, 86.8)
> cor(x,y)
[1] 0.768587
> cor(x,y,method="spearman")
```

```
[1] 1
> cor(x,y,method="kendall")
[1] 1
```

从 x 与 y 的散点图(见图4.32)可以看出, x 与 y 的线性相关系数受到右上角一个极端值的影而变小了. 因此在计算相关性度量时我们要考虑计算哪种相关系数更有意义.

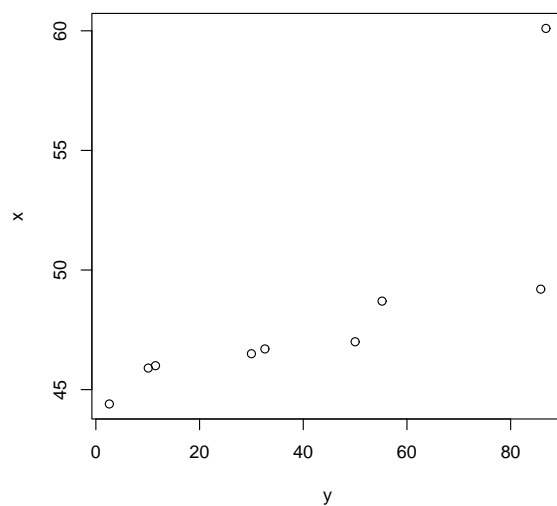


图 4.32 .

4.4.4 分组数据的图形概括

分组数据可视为特殊的多组数据, 他们的区别是: 在多组数据中各数值型变量的观测值指向不同的对象, 而分组数据是指同一个数值型变量的观测值按另一个分类变量分成若干个子集, 因此, 这些子集指向同一个变量. 下面我们通过DAAG中的数据集cuckoos来看一下分组数据的特殊图形描述方法.

例 4.4.5 杜鹃把蛋下在其它种类鸟的鸟巢中, 这些鸟会帮它们孵化, 我们希望了解在不同类的鸟巢中杜鹃蛋的长度, 数据如下:

```
> data(cuckoos)
```

```
> cuckoos
      length breadth      species id
1      21.7    16.1 meadow.pipit 21
2      22.6    17.0 meadow.pipit 22
... ...
118    20.8    15.9          wren 236
119    21.2    16.0          wren 237
120    21.0    16.0          wren 238
```

使用条件散点图

当数据集中含有一个或多个因子变量时, 可以使用条件散点图函数 `coplot()` 作出因子变量不同水平下的多个散点图, `coplot()` 的调用格式为:

—— `coplot()` 的调用格式 ——

```
> coplot(formula, data,.....)
```

对于一个因子变量 a , 变量 x 与 y 的条件散点图可用下面的命令得到:

```
> coplot(y ~ x | a)
```

对于二个因子变量 a 与 b , 变量 x 与 y 的条件散点图可用下面的命令得到:

```
> coplot(y ~ x | a*b)
```

对于例 4.4.5, 运行命令

```
> coplot(length ~ breadth | species)
```

得到图 4.33.

使用直方图

简单而繁琐的方法是反复使用函数 `hist()`. 运行命令

```
data(cuckoos)
attach(cuckoos)
```

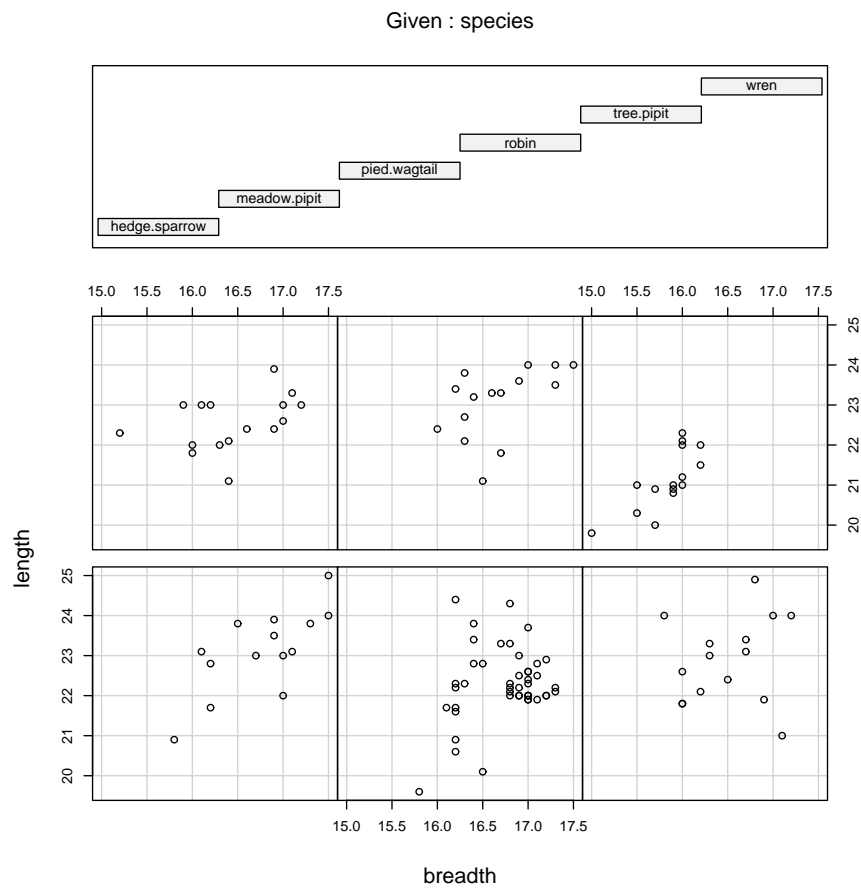


图 4.33 各鸟巢杜鹃蛋长度与宽度的散点图.

```

length.mp <- length[species=="meadow.pipit"]
length.tp <- length[species=="tree.pipit"]
length.hs <- length[species=="hedge.sparrow"]
length.r <- length[species=="robin"]
length.pw <- length[species=="pied.wagtail"]
length.w <- length[species=="wren"]
par(mfrow=c(3,2))
hist(length.mp,breaks=6,probability=T,
      xlim=c(19,25),ylim=c(0,1),main="",col=6)
hist(length.tp,breaks=6,probability=T,

```

```

xlim=c(19,25),ylim=c(0,1),main="",col=6)
hist(length.hs,breaks=6,probability=T,
      xlim=c(19,25),ylim=c(0,1),main="",col=6)
hist(length.r,breaks=6,probability=T,
      xlim=c(19,25),ylim=c(0,1),main="",col=6)
hist(length.pw,breaks=6,probability=T,
      xlim=c(19,25),ylim=c(0,1),main="",col=6)
hist(length.w,breaks=6,probability=T,
      xlim=c(19,25),ylim=c(0,1),main="",col=6)
par(mfrow=c(1,1))

```

得到图4.34.

我们可将上面的直方图纵向压缩在一起(像后面的框须图),得到所谓的直方组图. 直方组图函数**hists()**定义为:

直方组图函数**hists()**的定义

```

hists <- function (x, y, ...) {
  y <- factor(y)
  n <- length(levels(y))
  op <- par( mfcol=c(n,1), mar=c(2,4,1,1) )
  b <- hist(x, ..., plot=F)$breaks
  for (l in levels(y)){
    hist(x[y==l], breaks=b, probability=T, ylim=c(0,1.0),
         main="", ylab=l, col='lightblue', xlab="", ...)
    points(density(x[y==l]), type='l', lwd=3, col='red')
  }
  par(op)
}

```

ylim的范围可以根据需要自己调整,这更能直观地展示我们和数据. 由此运行命令

```
> hists(cuckoos$length,cuckoos$species)
```

运行得到图4.35.

我们也可直接利用**lattice**包中的直方图函数**histogram()**得到类似于4.34的每组数据的直方图. 运行命令

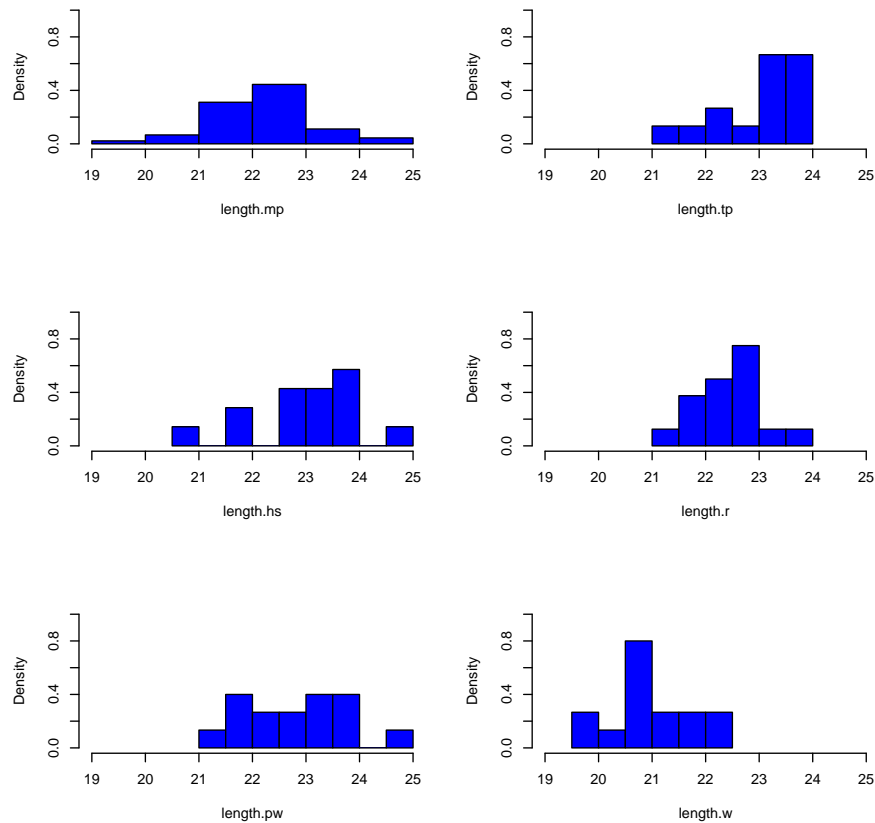


图 4.34 各鸟巢杜鹃蛋的直方图.

```
> histogram(~length|species,data=cuckoos)
```

到图4.36. 显然, 这种方法容易方便多了. `lattice`程序包还提供了其它许多功能强大、使用方便的其它作图函数, 有兴趣的读者可通过其帮助文件学习和使用.

使用框须图

我们可以用函数`boxplot()`同时考查各组数据的分布. 命令

```
> boxplot(length~species,data=cuckoos,
```

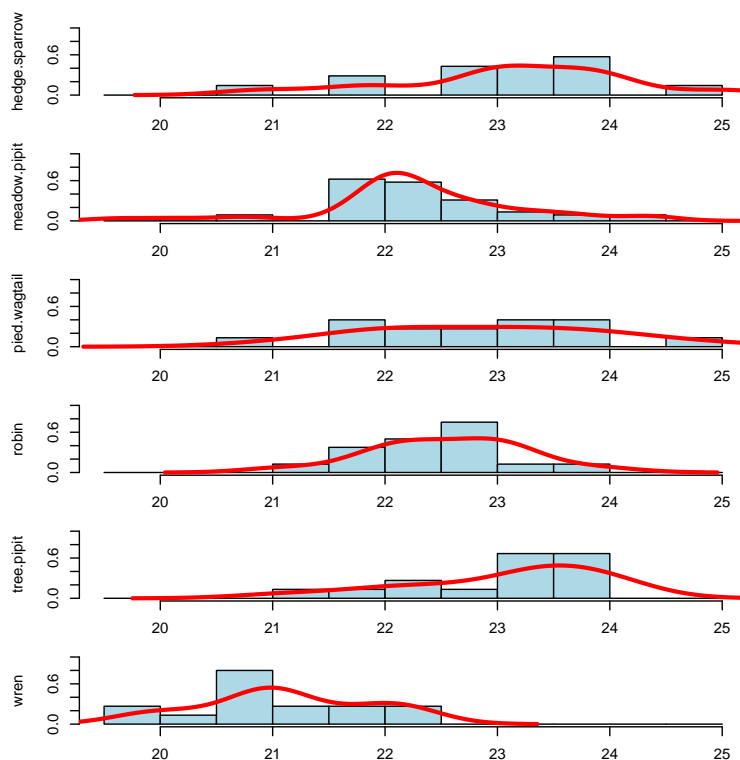


图 4.35 直方组图.

```
xlab="length of egg",horizontal=TRUE)
```

得到图4.37. 注意到horizontal=TRUE是让盒子横向放置. 从图上我们可以看出在wren(鸫)巢中的杜鹃蛋长度最小.

使用条形图

利用函数stripchart()得到杜鹃蛋在不同鸟巢的长度的分布图. 函数stripchart()在数据不多的时候和函数boxplot()的功能类似, 描绘了数据分布的情况, 其调用格式如下:

stripchart()的调用格式

```
> stripchart(x, method = "overplot"....)
```

说明: method说明数据重复的时候该如何放置, 有三种方式: overplot是重叠放置, stack是把数据垒起来, jitter是散放在数值的周围.

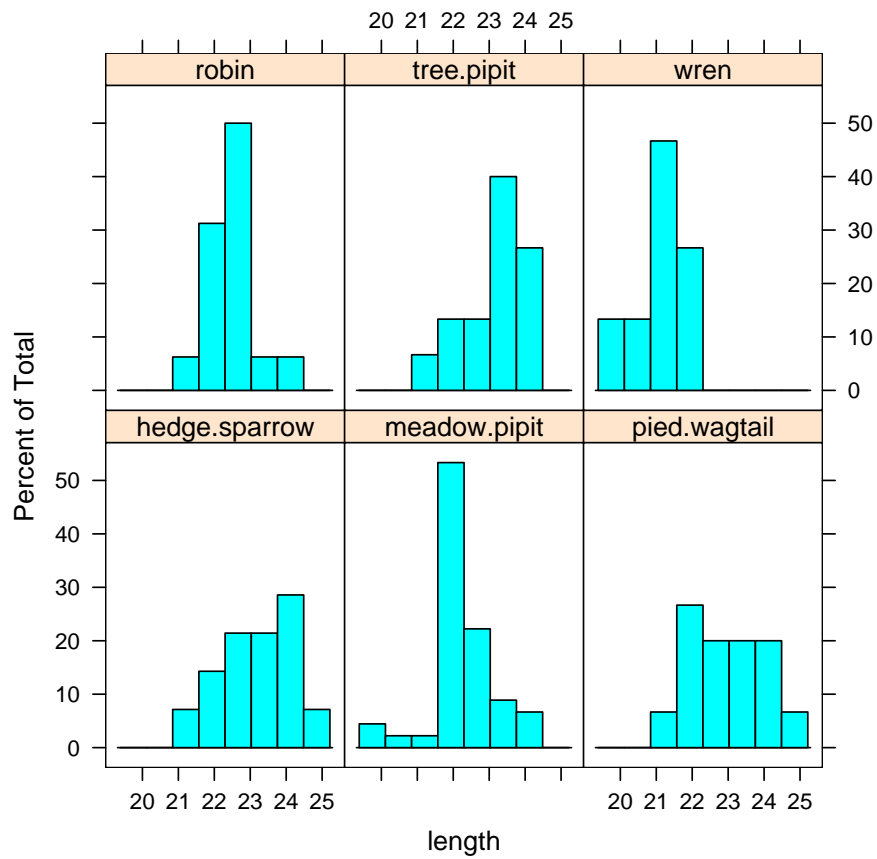


图 4.36 各鸟巢杜鹃蛋的直方图.

```
> stripchart(cuckoos$length~cuckoos$species, method ="jitter" )
```

运行得到图4.38.

使用密度曲线图

lattice包中的函数densityplot()可分别展示每组数据的密度曲线图.

```
> densityplot(~length|species,data=cuckoos)
```

运行得到图4.39.

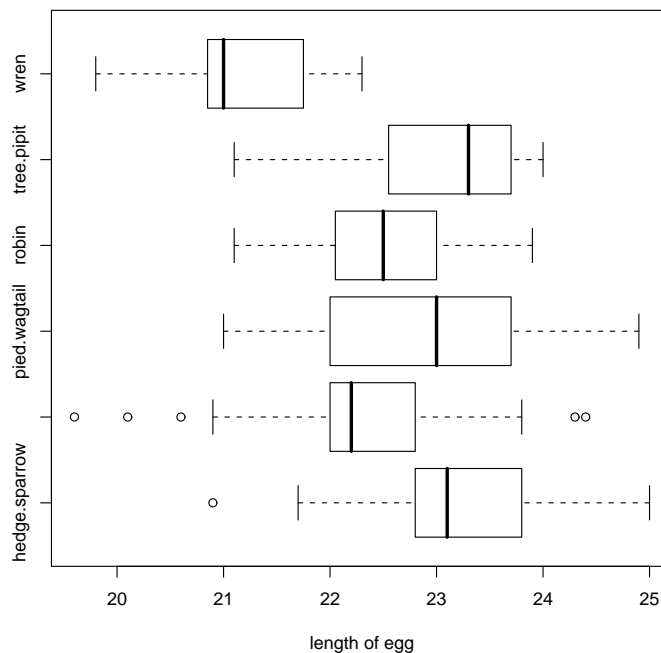


图 4.37 各鸟巢杜鹃蛋的boxplot图.

§4.5 分类数据的描述性统计分析

如果数据集中对应的变量都是定性变量, 这样的数据称为分类数据. 这种数据常使用表格来描述, 并为进一步的统计分析服务. 我们主要考虑由二元定性数据所构成的二维列联表数据. 这一节主要描述如何制作列联表和图形描述, 列联表的独立性检验将在第七章§7.3中介绍.

4.5.1 列联表的制作

由分类数据构造列联表

例 4.5.1 为考查眼睛的颜色(Eye)与头发的颜色(Hair)之间的关系, 收集了下面的一组数据

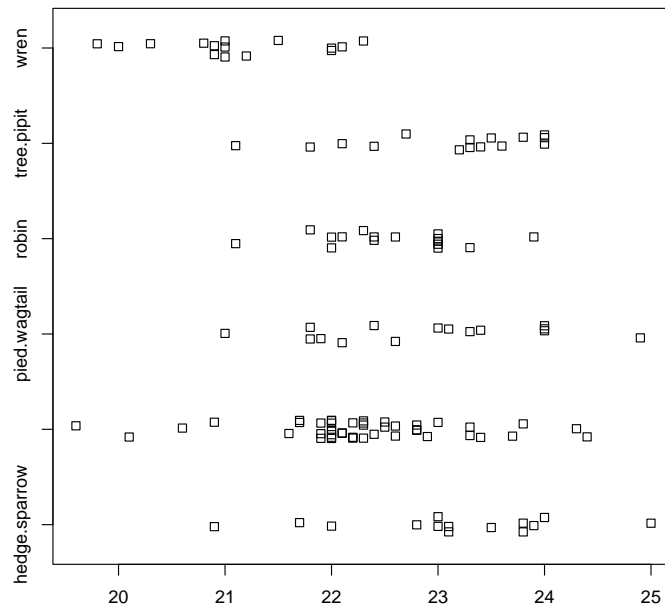


图 4.38 各鸟巢杜鹃蛋的stripchart图.

	Eye			
Hair	Brown	Blue	Hazel	Green
Black	68	20	15	5
Brown	119	84	54	29
Red	26	17	14	14
Blond	7	94	10	16

我们可以通过矩阵建立这个列联表, 命令如下

```
> Eye.Hair <- matrix(c(68,20,15,5, 119,84,54,29,
  26,17,14,14, 7,94,10,16), nrow=4,byrow=T)
> colnames(Eye.Hair) <- c("Brown", "Blue", "Hazel", "Green")
> rownames(Eye.Hair) <- c("Black","Brown","Red", "Blond")
> Eye.Hair
```

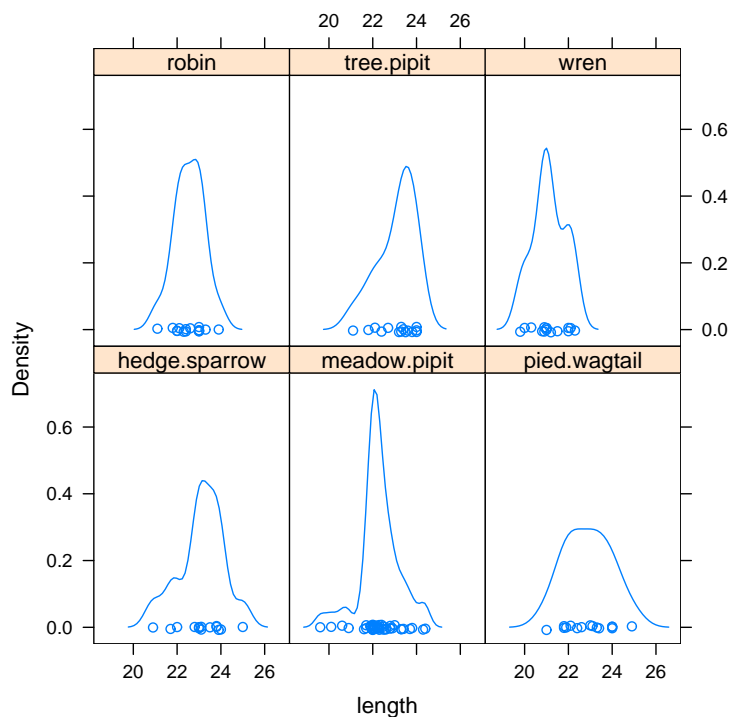


图 4.39 各鸟巢杜鹃蛋的密度曲线图。

由原始数据构造列联表

R中可以使用函数`table()`、`xtabs()`或`ftable()`由原始数据构造列联表, 具体用法参见它们的帮助. 我们仅以`table()`为例加以举例说明. 其用法为

`table()`的调用格式

```
> table(factor1, factor2, ...)
```

例 4.5.2 数据包ISwR中的数据集juul中含有三个分类变量: `sex`, `tanner`, `menarche`. 则我们可以得到下面的一些列表:

```
> table(sex)
> table(sex, menarche)
> table(menarche, tanner)
```

最后一个的显示结果为

```

      tanner
menarche  1   2   3   4   5
      1 221  43  32  14   2
      2   1   1   5  26 202

```

获得边际列表

在实际使用时常需要按列联表中某个属性(因子)求和, 称之为边际列表. 除了使用前面已经提到的函数`apply()`外, 更为方便的是使用函数`margin.table()`. 例如, 对于数据`Eye.Hair`, 我们有

```

> margin.table(Eye.Hair,1)
Black Brown   Red Blond
    108   286    71   127
> margin.table(Eye.Hair,2)
Brown Blue Hazel Green
    220   215    93    64

```

其中选项1和2分别表示按行和按列求边际和.

频率列联表

上面的列联表的元素为分类变量(因子)的频数, 故可称为频数列联表. 由频数列联表除以边际和就可得到它们的(相对)频率列联表, 这可通过函数`prop.table()`实现. 若再乘上100就得到相对应的用百分比表示的(相对)频率列联表. 仍以上面的例子加以说明

```

> prop.table(Eye.Hair,1)
      Brown Blue Hazel Green
Black  0.63  0.2  0.14  0.05
Brown  0.42  0.3  0.19  0.10
Red    0.37  0.2  0.20  0.20
Blond  0.06  0.7  0.08  0.13
> prop.table(Eye.Hair,1)*100
      Brown Blue Hazel Green

```

Black	63	19	14	5
Brown	42	29	19	10
Red	37	24	20	20
Blond	6	74	8	13

注意: 全局相对频率列联表不能由`prop.table()`得到, 但可以用下面的命令得到

```
> Eye.Hair/sum(Eye.Hair)
      Brown Blue Hazel Green
Black 0.11 0.03 0.03 0.008
Brown 0.20 0.14 0.09 0.049
Red   0.04 0.03 0.02 0.024
Blond 0.01 0.16 0.02 0.027
```

4.5.2 列联表的图形描述

使用条形图

像单组数据一样, 我们可以用条形图(或称为柱状图)来表示. 运行

```
> data(HairEyeColor)
> a <- as.table(apply(HairEyeColor,c(1,2),sum))
> barplot(a, legend.text = attr(a, "dimnames")$Hair)
```

得到图4.40. 这是按行(头发颜色)叠加、按列(眼睛颜色)排列的条形图. 我们也可将列并列放, 这时只需选项`beside`取值为`TRUE`. 运行

```
> barplot(a, beside = TRUE,
          legend.text = attr(a, "dimnames")$Hair)
```

得到图4.41.

使用点图

函数`dotchart()`给出Cleveland点图. 运行

```
> dotchart(Eye.Hair)
```

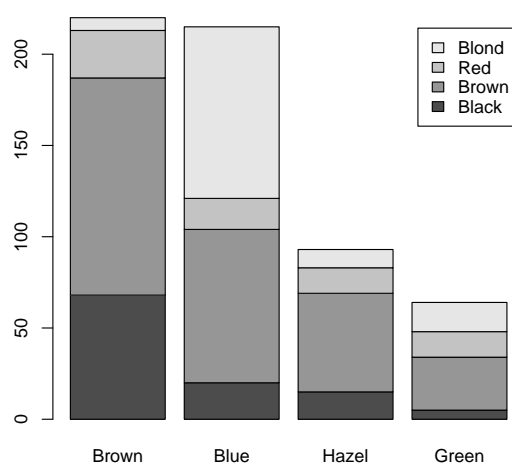


图 4.40 二元定性数据的条形图(1).

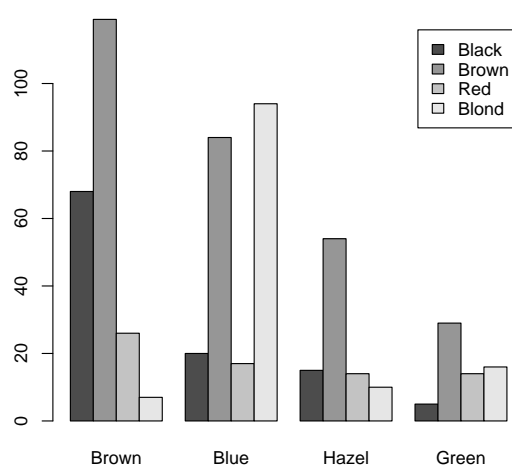


图 4.41 二元定性数据的条形图(2).

得到图4.42.

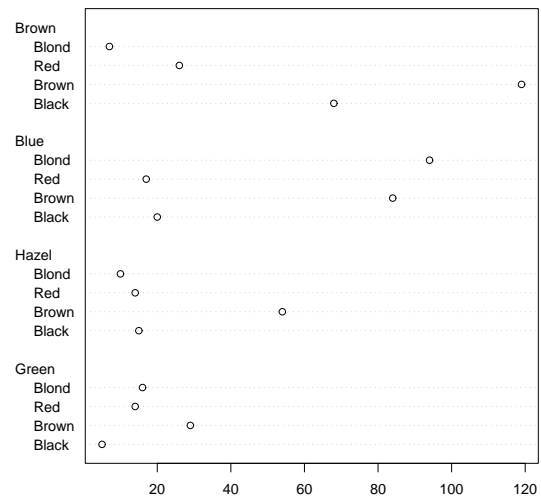


图 4.42 二元定性数据的Cleveland点图.

第四章习题

4.1 模拟得到1000个参数为0.3的贝努里分布随机数, 并用图示表示出来.

4.2 用命令`rnorm()`命令产生1000个均值为10, 方差为4的正态分布随机数, 用直方图呈现数据的分布并添加核密度曲线.

4.3 模拟得到三个t分布混合而成的样本, 用直方图呈现数据的分布并添加核密度曲线.

4.4 由程序包DAAG中的数据集聚sum,

- 1) 利用函数`hist(possum$age)`作出负鼠年龄的直方图. 试选用两种不同的断点并作比较, 说明两图的不同之处;
- 2) 求出负鼠年龄变量的均值、标准差、中位数以及上下四分位数.

4.5 考虑程序包DAAG中的数据集聚tinting,

- 1) 获得变量tint和sex的列联表;
- 2) 在同一图上作出变量sex与tint的联合柱状图;
- 3) 作出age和it的散点图, 并进一步完成下面的操作:
 - i. 用函数`lowness()`作出拟合线;
 - ii. 在图的两边加上更细小的刻度;
 - iii. 在图的两边加上箱型图.
- 4) 作出age和it关于因子变量tint的条件散点图;
- 5) 作出age和it关于因子变量tint和sex的条件散点图;
- 6) 做出it与csoa的等高线图;
- 7) 使用`matplot()`描述变量age, it和csoa.

4.6 由命令

```
> data(InsectSprays)
> InsectSprays
```

得到数据集 `InsectSprays`, 根据数据作出有意义的图, 并对数据作出描述性统计.

4.7 假定某校100名女生的血清总蛋白含量(g/L)服从均值为75, 标准差为3, 并假定数据由下面的命令产生

```
> options(digits=4)
> rnorm(100,75,9)
```

根据产生的数据

- 1) 计算样本均值、方差、标准差、极差、四分位极差、变异系数、偏度、峰度和五数概括;
- 2) 画出直方图、核密度估计曲线、经验分布图和QQ图;
- 3) 画出茎叶图、框须图.

4.8 某校测得20名学生的四项指标: 性别、年龄、身高(cm)和体重(kg), 具体数据如表4.1所示.

- 1) 绘制体重对身高的散点图;
- 2) 绘制不同性别下, 体重对身高的散点图;
- 3) 绘制不同年龄阶段, 体重对身高的散点图;
- 4) 绘制不同性别和不同年龄阶段, 体重对身高的散点图.

表 4.1: 学生身高与体重数据

学号	性别	年龄	身高	体重
01	F	18	166	54
02	F	18	155	58
03	F	19	154	50
04	F	18	160	47
05	F	20	162	46

(续下页)

学生身高与体重数据(续表)

学号	性别	年龄	身高	体重
06	F	19	153	48
07	F	21	156	50
08	F	20	152	49
09	F	21	170	57
10	F	20	156	52
11	M	18	168	61
12	M	18	166	55
13	M	19	172	63
14	M	18	178	68
15	M	20	169	59
16	M	19	180	65
17	M	21	177	59
18	M	20	168	56
19	M	21	182	69
20	M	20	170	61

第五章 参数估计

本章概要

- ◇ 矩法估计和极大似然估计
- ◇ 单正态总体的均值和方差的估计
- ◇ 两正态总体的参数估计
- ◇ 比率的估计
- ◇ 样本容量的确定

根据样本推断总体的分布和分布的数字特征称为统计推断. 这一章我们介绍统计推断的一个基本问题 — 参数估计问题. 在很多实际问题中, 总体的分布类型已知但它包含一个或多个参数, 总体的分布完全由所含的参数决定, 这样就需要对参数作出估计. 参数估计有两类, 一类是点估计, 就是以某个统计量的样本观测值作为未知参数的估计值; 另一类是区间估计, 就是用两个统计量所构成的区间来估计未知参数.

§5.1 矩法估计和极大似然估计

5.1.1 矩法估计

由辛钦大数定律和科尔莫哥洛夫强大数定理可知, 如果总体 X 的 k 阶矩存在, 则样本的 k 阶矩以概率收敛到总体的 k 阶矩, 样本矩的连续函数收敛到总体矩的连续函数. 这就启发我们可以用样本矩作为总体矩的估计量, 这种用相应的样本矩去估计总体矩的估计方法就称为矩估计法.

设 X_1, \dots, X_n 为来自某总体 X 的一个样本, 样本的 k 阶原点矩为

$$A_k = \frac{1}{n} \sum_{i=1}^n X_i^k, k = 1, 2, \dots$$

如果总体 X 的 k 阶原点矩 $\mu_k = E(X^k)$ 存在, 则按矩法估计的思想, 用 A_k 去估计 μ_k : $\hat{\mu}_k = A_k$.

设总体 X 的分布函数含有 k 个未知参数 $\theta = (\theta_1, \theta_2, \dots, \theta_k)$, 且分布的前 k 阶矩存在, 它们都是 $\theta_1, \theta_2, \dots, \theta_k$ 的函数, 此时求 $\theta_j (j = 1, 2, \dots, k)$ 的矩估计的具体步骤如下:

1) 求出 $E(X^j) = \mu_j, j = 1, 2, \dots, k$, 并假定

$$\mu_j = g_j(\theta_1, \theta_2, \dots, \theta_k), j = 1, 2, \dots, k. \quad (5-1.1)$$

2) 解方程组(5-1.1)得

$$\theta_i = h_i(\mu_1, \mu_2, \dots, \mu_k), i = 1, 2, \dots, k. \quad (5-1.2)$$

3) 在上式中用 A_j 代替 $\mu_j, j = 1, 2, \dots, k$ 即得 $\theta_1, \theta_2, \dots, \theta_k$ 的矩估计:

$$\hat{\theta}_i = h_i(A_1, A_2, \dots, A_k), i = 1, 2, \dots, k.$$

若有样本观测值 x_1, x_2, \dots, x_k , 代入上式即得 $\theta_1, \theta_2, \dots, \theta_k$ 的矩估计值.

由于函数 g_j 的表达式不同, 求解上述方程或方程组会相当困难, 这时需要通过迭代算法数值求解, 且这需要具体问题具体分析, 我们不可能有固定的 **R** 语言程序来直接估计 θ , 只能利用 **R** 的计算功能根据具体问题编写相应的 **R** 程序, 下面我们通过几个例子来说明如何在 **R** 中实现矩法估计.

例 5.1.1 设 X_1, \dots, X_n 是来自 $b(1, \theta)$ 的一个样本, θ 表示某件事件的成功概率, 通常事件的成功机会比 $g(\theta) = \theta/(1 - \theta)$ 是人们感兴趣的参数, 我们可以用矩法估计轻松地给出 $g(\theta)$ 一个很不错的估计, 因为 θ 是总体均值, 由矩法, 记 $\bar{X} = \frac{1}{n} \sum X_i$, 则

$$T(\bar{X}) = \frac{\bar{X}}{1 - \bar{X}}$$

是 $g(\theta)$ 的一个矩估计.

例 5.1.2 对某个篮球运动员记录其在一次比赛中投篮命中与否, 观测数据如下:

```
1 1 0 1 0 0 1 0 1 1 1 0 1 1 0 1
0 0 1 0 1 0 1 0 0 1 1 0 1 1 0 1
```

编写相应的R函数估计这个篮球运动员投篮的成败比.

```
> X<-c(1,1,0 ,1 ,0, 0, 1, 0 ,1 ,1,1, 0 ,1, 1 ,0 ,1,
        0 ,0 ,1, 0 ,1 ,0,1, 0 ,0 ,1,1 ,0 ,1, 1, 0, 1)
> theta<-mean(X)
> t<-theta/(1-theta)
> t
[1] 1.285714
```

我们得到 $g(\theta)$ 的矩估计为1.285714.

例 5.1.3 设总体为参数为 λ 的指数分布,其密度函数为

$$p(x|\lambda) = \lambda \exp^{-\lambda x}, \quad x > 0$$

X_1, \dots, X_n 是样本, 由于总体均值为 $1/\lambda$, 则 λ 的矩法估计为

$$\hat{\lambda} = \frac{1}{\bar{X}}$$

另外, 由于 $Var(X) = 1/\lambda^2$, 则 λ 的另一个矩法估计为

$$\hat{\lambda} = \frac{1}{\sqrt{s^2}}$$

其中 s^2 为样本方差. 这说明矩估计可能是不唯一的, 这是矩法估计的一个缺点, 此时通常应该尽量采用低阶矩给出未知参数的估计.

例 5.1.4 下面的观测值为来自指数分布的一个样本:

```
0.59132754  0.12854935  0.46900228  0.29835980  0.24341462
0.06566637  0.40085536  2.99687123  0.05278912  0.09898594
```

我们来估计其参数 λ .

R程序如下(一阶矩法估计):

```
> X<-c(0.59132754,0.12854935,0.46900228,0.29835980,0.24341462,
        0.06566637,0.40085536,2.99687123,0.05278912,0.09898594)
> lambda<- 1/mean(X)
> lambda
[1] 1.87062
```

如果使用二阶矩进行矩法估计, 则得

```
> lambda<- 1/sd(x)
> lambda
[1] 1.13103
```

结论:

- 1) λ 的一阶矩估计为1.87062, 二阶矩估计为1.13103. 实际上上面的数据是模拟参数为2的指数分布, 可见低阶矩更精确.
- 2) 在总体分布未知的情况下也可以用样本均值估计总体均值, 用样本方差估计总体方差.

5.1.2 极大似然估计

极大似然估计法是建立在极大似然原理基础上的一种统计方法, 我们先看一个例子: 某位同学与一位猎人一起外出打猎, 一只野兔从前方窜过. 只听一声枪响, 野兔应声倒下, 如果要你推测, 这一发命中的子弹是谁打的? 你就会想, 只发一枪便打中, 由于猎人命中的概率一般大于这位同学命中的概率, 看来这一枪是猎人射中的. 这种推断就体现了极大似然法的基本思想.

离散分布场合

设总体 X 是离散型随机变量, 其分布律为 $p(x|\theta)$, 其中 θ 是未知参数(或未知参数向量). 设 X_1, X_2, \dots, X_n 为取自总体 X 的样本, 则其联合概率函数为 $\prod_i^n p(x_i|\theta)$.

若我们已知样本的观测值为 x_1, x_2, \dots, x_n , 则事件 $(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$ 发生的概率为 $\prod_i^n p(x_i|\theta)$. 这一概率随 θ 的值而变化. 从直观上来看, 既然样本观测值 x_1, x_2, \dots, x_n 出现了, 它们出现的概率, 即 $\prod_i^n p(x_i|\theta)$ 相

对来说应比较大. 换句话说, θ 应使样本 x_1, x_2, \dots, x_n 的出现具有较大的概率. 将上式看作 θ 的函数, 并用 $L(\theta)$ 表示, 即

$$L(\theta) = L(x_1, x_2, \dots, x_n; \theta) = \prod_i^n p(x_i | \theta). \quad (5-1.3)$$

称 $L(\theta)$ 为似然函数. 极大似然估计法就是在参数 θ 的可能取值范围 Θ 内, 选取使 $L(\theta)$ 达到最大的参数值 $\hat{\theta}$ 作为参数 θ 的估计值. 即取 $\hat{\theta}$, 使

$$L(\theta) = L(x_1, x_2, \dots, x_n; \hat{\theta}) = \max_{\theta \in \Theta} L(x_1, x_2, \dots, x_n; \theta).$$

连续分布场合

设总体 X 是连续型随机变量, 其概率密度函数为 $p(x|\theta)$, 其中 θ 是未知参数(或未知参数向量). 设 X_1, X_2, \dots, X_n 为取自总体 X 的样本, 则其联合密度函数值为 $\prod_i^n f(x_i|\theta)$. 若取得样本观察值为 x_1, x_2, \dots, x_n , 则因为 (X_1, X_2, \dots, X_n) 取 (x_1, x_2, \dots, x_n) (指落在其邻域中) 的概率正比于 $\prod_i^n p(x_i|\theta)$, 所以, 按极大似然原理, 应选择 θ 的值使此概率达到最大. 我们也称 $L(\theta) = \prod_i^n f(x_i|\theta)$ 为似然函数. 再按离散场合同样的方法求使似然函数达到最大的参数 θ 的值, 即极大似然估计值.

可见, 不管在离散还是连续场合, 似然函数都可表示为(5-1.3), 其中 $p(x|\theta)$ 为总体 X 的概率函数, 它在离散表示分布律, 在连续场合表示密度函数.

在单参数场合, 我们可以使用 **R** 中的函数 `optimize()` 求极大似然估计值. `optimize()` 的调用格式如下:

optimize() 的调用格式

```
optimiz(f = , interval = , lower = min(interval),
        upper = max(interval), maximum = TRUE,
        tol = .Machine$double.eps^0.25, ...)
```

说明: `f` 是似然函数, `interval` 是参数 θ 的取值范围, `lower` 是 θ 的下界, `upper` 是 θ 的上界, `maximum = TRUE` 是求极大值, 否则(`maximum = FALSE`)表示求函数的极小值, `tol` 是表示求值的精确度, ... 是对 `f` 的附加说明.

在多参数场合, 我们用函数 `optim()` 或者 `nlm()` 来求似然函数的极大值, 并求相应的极大值点. `optim()` 和 `nlm()` 的定义如下:

—— `optim()` 的调用格式 ——

```
optim(par, fn, gr = NULL,
      method = c("Nelder-Mead", "BFGS", "CG", "L-BFGS-B", "SANN"),
      lower = -Inf, upper = Inf,
      control = list(), hessian = FALSE, ...)
```

—— `nlm()` 的调用格式 ——

```
nlm(f, p, hessian = FALSE, typsize=rep(1, length(p)), fscale=1,
    print.level = 0, ndigit=12, gradtol = 1e-6,
    stepmax = max(1000 * sqrt(sum((p/typsize)^2)), 1000),
    steptol = 1e-6, iterlim = 100, check.analyticals = TRUE, ...)
```

三者的主要区别是：函数`nlm()`仅使用牛顿-拉夫逊算法求函数的最小值点；函数`optim()`提供`method`选项给出的5种方法中的一种进行优化；上面二个可用于多维函数的极值问题，而函数`optimize()`仅适用于一维函数，但可以用于最大与最小值点。

下面通过一个例子来说明 θ 为一维时如何求极大似然估计。

例 5.1.5 一地质学家为研究密歇根湖的糊滩地区的岩石成分，随机地自该地区取出100个样品，每个样品有十块石子，他记录了每个样品中属石灰石的石子数，所得到的数据如表5.1所示。假设这100次观测相互独立，求这地区石子中的石灰石的比例 p 的最大似然估计。

表 5.1 岩石成分数据

样本中的石子数	0	1	2	3	4	5	6	7	8	9	10
样品个数	0	1	6	7	23	26	21	12	3	1	2

解 显然，每个样品中的石子数服从二项分布 $b(10, p)$ ，我们的目的是根据100次观测估计参数 p 。似然函数为

$$\begin{aligned}
 L(\theta) &= L(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n p(x_i, \theta) \\
 &= p^{1+2 \times 6 + \dots + 10 \times 2} (1-p)^{100 \times 10 - (1+2 \times 6 + \dots + 10 \times 2)} \\
 &= p^{517} (1-p)^{483}
 \end{aligned}$$

R中程序如下：

```
> f <- function(P) (P^517)*(1-P)^483
> optimize(f,c(0,1),maximum = TRUE)
$maximum
[1] 0.5170006
$objective
[1] 1.663700e-301
```

因此该地区石子中的石灰石的比例 p 的最大似然估计为0.517. 在计算结果中, `$maximum`是极大值的近似解, 即估计值 $\hat{p} = 0.5170$, `$objective`是目标函数在近似解处的函数值. ■

§5.2 单正态总体参数的区间估计

上一节我们讨论了点估计, 由于点估计值只是估计量的一个近似值, 因而点估计本身既没有反映出这种近似值的精度, 即指出用估计值去估计的误差范围有多大, 而且也没有指出这个误差范围以多大的概率包括未知参数, 这些问题正是区间估计要讨论的问题. 区间估计解决了这二个问题, 它给出了估计的可信程度, 是一种重要的统计推断形式. 我们在接下来的几节将讨论这个问题. 这一节我们讨论单正态总体参数的区间估计问题.

假设总体 $X \sim N(\mu, \sigma^2)$, X_1, \dots, X_n 是来自此正态总体的一个样本, $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 为其样本均值, $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 为其样本方差.

5.2.1 均值 μ 的区间估计

1. 方差 σ^2 已知时 μ 的置信区间

由于

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

因此有

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1). \quad (5-2.1)$$

由 $P(-z_{1-\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}) = 1 - \alpha$ 即得

$$P\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} < \mu < \bar{X} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

所以, 对于单个正态母体 $N(\mu, \sigma^2)$, 在 σ^2 已知时, μ 的置信度为 $1 - \alpha$ 的置信区间为

$$\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \right),$$

简记为

$$\bar{X} \pm \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}.$$

同理可求得 μ 的置信度为 $1 - \alpha$ 的单侧置信上限

$$\bar{X} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha},$$

μ 的置信度为 $1 - \alpha$ 的单侧置信下限为

$$\bar{X} - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}.$$

由在 \mathbf{R} 中没有求方差已知时均值置信区间的内置函数, 需要自己编写函数. 编写的 \mathbf{R} 程序如下:

z.test()函数的定义

```
z.test<-function(x,n,sigma,alpha,u0=0,alternative="two.sided"){
  options(digits=4)
  result<-list( )
  mean<-mean(x)
  z<-(mean-u0)/(sigma/sqrt(n))
  p<-pnorm(z,lower.tail=FALSE)
  result$mean<-mean
  result$z<-z
  result$p.value<-p
  if(alternative=="two.sided"){
    p<-2*p
    result$p.value<-p
  }
  else if (alternative == "greater"|alternative == "less" ){
    result$p.value<-p
  }
  else return("your input is wrong")
  result$conf.int<- c(
    mean-sigma*qnorm(1-alpha/2,mean=0, sd=1,
      lower.tail = TRUE)/sqrt(n),
```

```

        mean+sigma*qnorm(1-alpha/2,mean=0, sd=1,
                        lower.tail = TRUE)/sqrt(n))
    result
}

```

利用此程序即可给出体均值的置信区间. 此程序还可用于进行第七章要讲的正态总体均值 μ 的假设检验, 之所以在程序中同时完成区间估计与假设检验, 是为了与R中的t检验函数`t.test()`相对应. 实际上, 我们可以从上面的程序中抽出区间估计的部分, 得到下面求置信区间的程序:

```

> conf.int<-function(x,n,sigma,alpha){
  options(digits=4)
  mean<-mean(x)
  c(mean-sigma*qnorm(1-alpha/2,mean=0, sd=1,
                    lower.tail = TRUE)/sqrt(n),
    mean+sigma*qnorm(1-alpha/2,mean=0, sd=1,
                    lower.tail = TRUE)/sqrt(n))
}

```

下面通过例子看一下在R中如何去求置信度为 $1 - \alpha$ 的置信区间.

例 5.2.1 一个人10次称自己的体重(单位:斤): 175 176 173 175 174 173 173 176 173 179, 我们希望估计一下他的体重. 假设此人的体重服从正态分布, 标准差为1.5, 我们要求体重的置信水平为95%的置信区间.

解 由上述函数`z.test()`, R程序为

```

> x<-c(175 ,176 ,173,175,174,173,173,176,173,179 )
> result<-z.test(x,10,1.5,0.05)
> result$conf.int
[1] 173.8 175.6

```

因此, 我们得到体重的置信水平为0.95的置信区间为(173.8 ,175.6).

注: 运行

```

> z.test(x, 10, 1,5, 0.05)

```

将同时获得假设检验的结果, 而上面的程序仅提取了区间估计的部分, 这相当于执行了

```
> x<-c(175 ,176 ,173,175,174,173,176,173,179 )
> conf.int(x,10,1.5,0.05)
```

■

2. 方差 σ^2 未知时 μ 的置信区间

由于

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1),$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

且二者独立, 所以有

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1). \quad (5-2.2)$$

同样由 $P(-t_{1-\frac{\alpha}{2}}(n-1) < T < t_{1-\frac{\alpha}{2}}(n-1)) = 1 - \alpha$ 得到

$$P\left(\bar{X} - \frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1) < \mu < \bar{X} + \frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1)\right) = 1 - \alpha,$$

所以, 在 σ^2 未知时, μ 的置信度为 $1 - \alpha$ 的置信区间为

$$\left(\bar{X} - \frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1), \bar{X} + \frac{S}{\sqrt{n}}t_{1-\frac{\alpha}{2}}(n-1)\right),$$

其中 $t_p(n)$ 为自由度为 n 的 t 分布的下侧 p 分位数. 同理可求得 μ 的置信度为 $1 - \alpha$ 的单侧置信上限为

$$\bar{X} + \frac{S}{\sqrt{n}}t_{1-\alpha}(n-1),$$

μ 的置信度为 $1 - \alpha$ 的单侧置信下限为

$$\bar{X} - \frac{S}{\sqrt{n}}t_{1-\alpha}(n-1).$$

方差未知时我们直接利用R语言的`t.test()`来求置信区间. `t.test()`的调用格式如下:

`t.test()`的调用格式

```
t.test(x, y = NULL,
```

```
alternative = c("two.sided", "less", "greater"),  
mu = 0, paired = FALSE, var.equal = FALSE,  
conf.level = 0.95, ...)
```

说明: 若仅出现数据 x , 则进行单样本 t 检验; 若出现数据 x 和 y , 则进行二样本的 t 检验(见6.3节); `alternative=c("two.sided", "less", "greater")`用于指定所求置信区间的类型; `alternative="two.sided"`是缺省值, 表示求置信区间;`alternative="less"`表示求置信上限; `alternative="greater"`表示求置信下限. `mu`表示均值, 它仅在假设检验中起作用, 默认值为零.

在上例中如果不知道方差, 就需要用函数`t.test()`来求置信区间, 我们看一下在R中是如何实现的.

R程序如下:

```
> x<-c(175 , 176 , 173 , 175 ,174 ,173 , 173, 176 , 173,179 )  
> t.test(x)
```

运行结果如下:

```
One Sample t-test  
data:  x  
t = 283.8161, df = 9, p-value < 2.2e-16  
alternative hypothesis: true mean is not equal to 0  
95 percent confidence interval:  
 173.3076 176.0924  
sample estimates:  
mean of x  
 174.7
```

我们可以看到置信水平为0.95的置信区间为(173.3076, 176.0924).

我们注意到这个输出结果过于繁琐, 关于假设检验的结果仅在第六章中用到. 由于我们只需要置信区间的结果, 因此R程序:

```
> t.test(x)$conf.int
```

提取出置信区间的部分, 结果如下:

```
[1] 173.3076 176.0924
attr(,"conf.level")
[1] 0.95
```

以下用到的许多程序都可能输出很多结果,如同此例,在其后面加上\$conf.int将只输出置信区间的结果.

5.2.2 方差 σ^2 的区间估计

此时虽然也可以就均值是否已知分两种情况讨论 σ^2 的置信区间,但在实际中 μ 已知的情形是极为罕见的,所以我们只在 μ 未知的条件下讨论 σ^2 的置信区间.

由于

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1), \quad (5-2.3)$$

所以由

$$P\left(\chi_{\frac{\alpha}{2}}^2(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi_{1-\frac{\alpha}{2}}^2(n-1)\right) = 1 - \alpha$$

就可得到 σ^2 的置信水平为 $1 - \alpha$ 的置信区间

$$\left(\frac{(n-1)S^2}{\chi_{1-\frac{\alpha}{2}}^2(n-1)}, \frac{(n-1)S^2}{\chi_{\frac{\alpha}{2}}^2(n-1)} \right).$$

在R中也没有直接求 σ^2 的置信区间的函数,我们需要编写自己需要的函数,下面的函数chisq.var.test()可以用来求 σ^2 置信区间.(第六章还将用于关于 σ^2 的假设检验.)

chisq.var.test()的定义

```
chisq.var.test <- function(x,var,alpha,alternative="two.sided"){
  options(digits=4)
  result<-list( )
  n<-length(x)
  v<-var(x)
  result$var<-v
  chi2<-(n-1)*v/var
  result$chi2<-chi2
  p<-pchisq(chi2,n-1)
```

```

if(alternative == "less"|alternative=="greater"){
  result$p.value<-p
} else if (alternative=="two.sided") {
  if(p>.5)
    p<-1-p
  p<-2*p
  result$p.value<-p
} else return("your input is wrong")
result$conf.int<-c(
  (n-1)*v/qchisq(alpha/2, df=n-1, lower.tail=F),
  (n-1)*v/qchisq(alpha/2, df=n-1, lower.tail=T))
result
}

```

将此函数用到上例, 由运行显示 σ^2 的0.95置信区间为(1.793, 12.628).

§5.3 两正态总体参数的区间估计

设总体 X 与 Y 独立, $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, X_1, \dots, X_{n_1} 是来自总体 X 的样本, $\bar{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i$ 为其样本均值, $S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$ 为其样本方差. Y_1, \dots, Y_{n_2} 是来自总体 Y 的样本, $\bar{Y} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i$ 为其样本均值, $S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$ 为其样本方差.

5.3.1 均值差 $\mu_1 - \mu_2$ 的置信区间

1. 两方差都已知时两均值差的置信区间

进一步假设 σ_1^2 与 σ_2^2 都已知, 要求 $\mu_1 - \mu_2$ 置信水平为 $1 - \alpha$ 的置信区间. 由于

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \quad \bar{Y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right),$$

且两者独立, 得

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right),$$

所以

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1). \quad (5-3.1)$$

由

$$P(-z_{1-\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}) = 1 - \alpha,$$

化简得

$$P\left(\bar{X} - \bar{Y} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \bar{X} - \bar{Y} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha.$$

所以 $\mu_1 - \mu_2$ 的置信水平 $1 - \alpha$ 的置信区间为

$$\left(\bar{X} - \bar{Y} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \quad \bar{X} - \bar{Y} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

同理可求得 $\mu_1 - \mu_2$ 的置信水平 $1 - \alpha$ 的单侧置信上限为

$$\bar{X} - \bar{Y} + z_{1-\alpha} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

$\mu_1 - \mu_2$ 的置信水平 $1 - \alpha$ 的单侧置信下限为

$$\bar{X} - \bar{Y} - z_{1-\alpha} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

在R语言中可以编写函数求置信区间(单侧置信限读者可以类似地编写程序)

```

two.sample.ci( )的定义
two.sample.ci<-function(x,y,conf.level=0.95, sigma1,sigma2 ){
  options(digits=4)
  m= length(x); n = length(y)
  xbar=mean(x)-mean(y) alpha = 1 - conf.level
  zstar= qnorm(1-alpha/2)* (sigma1/m+sigma2/n)^(1/2)
  xbar +c(-zstar, +zstar)
}

```

我们来看一个例子.

例 5.3.1 为比较两个小麦品种的产量, 选择18块条件相似的试验田, 采用相同的耕作方法做实验, 结果播种甲品种的8块实验田的单位面积产量和播种乙品种的10块实验田的单位面积产量分别为:

甲品种	628	583	510	554	612	523	530	615		
乙品种	535	433	398	470	567	480	498	560	503	426

假定每个品种的单位面积产量均服从正态分布, 甲品种产量的方差为2140, 乙品种产量的方差为3250, 试求这两个品种平均面积产量差的置信区间(取 $\alpha=0.05$)

解 直接利用上面编写的函数:

```
> x<-c(628,583,510,554,612,523,530,615)
> y<-c(535,433,398,470,567,480,498,560,503,426)
> sigma1<-2140
> sigma2<-3250
> two.sample.ci(x,y,conf.level=0.95, sigma1,sigma2)
[1] 34.67 130.08
```

所以这两个品种平均面积产量差的置信水平0.95的置信区间为(34.67, 130.08).

■

2. 两方差都未知时两均值差的置信区间

设方差 σ_1^2 与 σ_2^2 都未知, 但 $\sigma_1^2 = \sigma_2^2 = \sigma^2$. 此时由于

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma^2 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1),$$

$$\frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2(n_1 - 1), \quad \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2(n_2 - 1)$$

且由 S_1^2 与 S_2^2 的相互性得

$$\frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2).$$

由此可以得到

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})S^2}} \sim t(n_1 + n_2 - 2), \quad (5-3.2)$$

其中

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}.$$

由

$$P(-t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2) < T < t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2)) = 1 - \alpha$$

解不等式即得 $\mu_1 - \mu_2$ 的置信水平为 $1 - \alpha$ 的置信区间:

$$\bar{X} - \bar{Y} \pm t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2) \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} S.$$

同理可求得 $\mu_1 - \mu_2$ 的置信水平为 $1 - \alpha$ 的单侧置信上限为

$$\bar{X} - \bar{Y} + t_{1-\alpha}(n_1 + n_2 - 2) \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} S,$$

$\mu_1 - \mu_2$ 的置信水平为 $1 - \alpha$ 的单侧置信下限为

$$\bar{X} - \bar{Y} - t_{1-\alpha}(n_1 + n_2 - 2) \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} S.$$

如同求单正态总体的均值的置信区间, 在R中可以直接利用`t.test()`求两方差都未知但相等时两均值差的置信区间.

例 5.3.2 在例5.3.1中, 如果不知道两种品种产量的方差但已知两者相等, 此时须在`t.test()`中指定选项`var.equal=TRUE`, 则由

```
> x<-c(628,583,510,554,612,523,530,615)
> y<-c(535,433,398,470,567,480,498,560,503,426)
> t.test(x,y,var.equal=TRUE)
```

运行得到

```
Two Sample t-test
data:  x and y
```

```
t = 3.3007, df = 16, p-value = 0.004512
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 29.46961 135.28039
sample estimates:
mean of x mean of y
 569.375   487.000
```

可见, 这两个品种的单位面积产量之差的置信水平0.95的置信区间为(29.4696, 135.2804).

5.3.2 两方差比 σ_1^2/σ_2^2 的置信区间

由于

$$\frac{(n_1 - 1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1), \quad \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1),$$

且 s_1^2 与 s_2^2 相互独立, 故

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1). \quad (5-3.3)$$

所以, 对给定的置信水平 $1 - \alpha$, 由

$$P\left(F_{\alpha/2}(n_1 - 1, n_2 - 1) \leq \frac{S_1^2}{S_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} \leq F_{1-\alpha/2}(n_1 - 1, n_2 - 1)\right) = 1 - \alpha,$$

经不等式变形即得 σ_1^2/σ_2^2 的 $1 - \alpha$ 置信区间

$$\left(\frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{1-\alpha/2}(n_1 - 1, n_2 - 1)}, \quad \frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{\alpha/2}(n_1 - 1, n_2 - 1)}\right),$$

其中 $F_p(m, n)$ 为自由度为 (m, n) 的 F 分布的下侧 p 分位数.

R中函数`var.test()`可以直接用于求两正态总体方差比的置信区间, 其调用格式如下:

var.test() 的调用格式

```
var.test(x, y, ratio = 1,
         alternative = c("two.sided", "less", "greater"),
```

```
conf.level = 0.95, ...)
```

在求置信区间时, 我们只需给出两个总体的样本 x, y 以及相应的置信水平, 选项`alternative`用于下一章的假设检验. 我们用下面的例子来说明.

例 5.3.3 甲、乙两台机床分别加工某种轴承, 轴承的直径分别服从正态分布 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$, 从各自加工的轴承中分别抽取若干个轴承测其直径, 结果如表5.2所示. 试求两台机床加工的轴承直径的方差比 σ_1^2/σ_2^2 的0.95置信区间.

表 5.2 机床加工的轴的直径数据

总体	样本容量	直径
X(机床甲)	8	20.5 19.8 19.7 20.4 20.1 20.0 19.0 19.9
Y(机床乙)	7	20.7 19.8 19.5 20.8 20.4 19.6 20.2

解 R程序如下:

```
> x<-c(20.5,19.8,19.7,20.4,20.1,20.0,19.0,19.9)
> y<-c(20.7,19.8,19.5,20.8,20.4,19.6,20.2)
> var.test(x,y)
```

运行结果如下:

```
F test to compare two variances

data:  x and y
F = 0.7932, num df = 7, denom df = 6, p-value = 0.7608
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1392675 4.0600387
sample estimates:
ratio of variances
 0.7931937
```

可见两台机床的加工的轴承的直径的方差比 σ_1^2/σ_2^2 的0.95置信区间为(0.1393, 4.0600). 结果中`sample estimates`给出的是方差比 σ_1^2/σ_2^2 的矩估计

值0.7931937. ■

§5.4 单总体比率 p 的区间估计

在许多实际问题中, 我们经常要去估计在总体中具有某种特性的个体占总体的比例(率), 设为 p . 例如, 整个学校中女生(或男生)占全校人数的比例, 一批产品中合格产品占总产品数的比例, 产品的不合格品率、某一电视节目的收视率、对某项政策的支持率等等. 关于点估计我们在第一节已经介绍, 这里介绍一种求 p 的近似区间估计的方法.

称在样本中具有某种特征的个体占样本总数的比例为样本比例. 设 x 为容量为 n 的样本中具有某种特征的个体数量, 则样本比例为 x/n . 当总体中的样品数足够多时, x 近似服从二项分布 $b(n, p)$ (实际上它是超几何分布), 这时总体比例可用样本比例来估计, 即 $\hat{p} = \frac{x}{n}$, 且为极大似然估计. 当 n 较大时, 由中心极限定理知 \hat{p} 具有渐近正态性, 即

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0, 1).$$

由于 n 较大, 所以可用 \hat{p} 来代替分母中的 p , 从而近似地有

$$Z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim N(0, 1). \quad (5.4.1)$$

这样由

$$P(-z_{1-\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

解不等式即得总体比例 p 的置信度为 $1 - \alpha$ 的置信区间

$$\left(\hat{p} - z_{1-\frac{\alpha}{2}} \sqrt{\hat{p}(1-\hat{p})/n}, \quad \hat{p} + z_{1-\frac{\alpha}{2}} \sqrt{\hat{p}(1-\hat{p})/n} \right)$$

同理可得 p 的置信度为 $1 - \alpha$ 的单侧置信上限为

$$\hat{p} + z_{1-\alpha} \sqrt{\hat{p}(1-\hat{p})/n},$$

p 的置信度为 $1 - \alpha$ 的单侧置信下限为

$$\hat{p} - z_{1-\alpha} \sqrt{\hat{p}(1-\hat{p})/n}.$$

在R中, 我们可直接利用函数`prop.test()`对 p 进行估计与检验, 其调用格式如下:

—— `prop.test()` 的调用格式 ——

```
prop.test(x, n, p = NULL,
          alternative = c("two.sided", "less", "greater"),
          conf.level = 0.95, correct = TRUE)
```

说明: x 为样本中具有某种特性的样本数量, n 为样本容量, `correct`选项为是否做连续性校正. 根据抽样理论, p 的 $1 - \alpha$ 的近似置信区间为

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{(1-f)\hat{p}(1-\hat{p})/(n-1)} - \frac{1}{2n},$$

其中 f 为抽样比. 由于假设样本容量很大, 因此修正后 p 的置信度为 $1 - \alpha$ 的置信区间近似地为

$$\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\hat{p}(1-\hat{p})/n} - \frac{1}{2n}.$$

它与刚才用中心极限定理推得的结论相比, 区间长了 $\frac{1}{n}$, 这是由于用连续分布去近似离散分布(超几何分布)引起的.

例 5.4.1 从一份共有3042人的人名录中随机抽200人, 发现38人的地址已变动, 试以95%的置信度, 估计这份名录中需要修改地址的比例.

解 在R中键入

```
> prop.test(38,200,correct=TRUE)
```

得到如下的结果:

```
1-sample proportions test with continuity correction
data: 38 out of 200, null probability 0.5
X-squared = 75.645, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.1394851 0.2527281
sample estimates:
p
0.19
```

所以我们以95%的置信水平认为这份名录中需要修改地址的比例 p 为落在(0.1395, 0.2527)中, 其点估计为0.19.

如果不进行校正,相应的R语句为:

```
> prop.test(38,200,correct=FALSE)
```

结果如下:

```
1-sample proportions test without continuity correction
data: 38 out of 200, null probability 0.5
X-squared = 76.88, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.1416717 0.2500124
sample estimates:
p
0.19
```

此时 p 的95%置信区间为(0.1417, 0.2500), 其长度比修正的缩短了. ■

前已指出, 样本中具有某种特性的样本数量 x 服从超几何分布, 上面我们用正态分布来近似, 还可以用二项分布来近似超几何分布, 此时要求抽样比 f 很小. R中函数`binom.test()`可以求其置信区间, 其调用格式如下:

binom.test() 的调用格式

```
binom.test(x, n, p = NULL,
           alternative = c("two.sided", "less", "greater"),
           conf.level = 0.95)
```

其含义和上面的函数`prop.test()`一致. 用到上例中, 由

```
> binom.test(38,200)
```

得结果如下:

```
Exact binomial test
data: 38 and 200
number of successes = 38, number of trials = 200, p-value < 2.2e-16
```


alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:

0.1381031 0.2513315

sample estimates:

probability of success

0.19

可见用二项分布近似所得的 p 的95%置信区间为(0.1381, 0.2513), 它与修正的正态近似方法更接近.

§5.5 两总体比率差 $p_1 - p_2$ 的区间估计

设有两总体 X 与 Y 相互独立(总体容量都较大), 从中分别抽取 n_1 和 n_2 个(n_1, n_2 也较大)观察, 结果发现其中各有 x_1 和 x_2 个具有某种特性. 设总体 X 与 Y 中具有上述待性的比率分别为 p_1 和 p_2 , 我们的目的是要估计 $p_1 - p_2$, 我们仅考虑近似正态性下的区间估计问题.

两个总体比例 p_1 和 p_2 的极大似然估计分别为 $\hat{p}_1 = \frac{x_1}{n_1}, \hat{p}_2 = \frac{x_2}{n_2}$. 由上一节, 若 n_1 和 n_2 较大, 则 \hat{p}_1, \hat{p}_2 近似地服从正态分布:

$$\hat{p}_1 \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right), \quad \hat{p}_2 \sim N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right).$$

所以

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right).$$

标准化, 并用 \hat{p}_1, \hat{p}_2 分别代替 p_1, p_2 , 得到

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim N(0, 1). \quad (5-5.1)$$

这样由

$$P(-z_{1-\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}) = 1 - \alpha,$$

通过不等式变形即得两比例差 $p_1 - p_2$ 的置信水平为 $1 - \alpha$ 的区间估计:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

同理可得 $p_1 - p_2$ 的置信水平为 $1 - \alpha$ 的单侧置信上限为

$$(\hat{p}_1 - \hat{p}_2) + z_{1-\alpha} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

$p_1 - p_2$ 的置信水平为 $1 - \alpha$ 的单侧下限为

$$(\hat{p}_1 - \hat{p}_2) - z_{1-\alpha} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

例 5.5.1 据一项市场调查, 在A地区被调查的1000人中有478人喜欢品牌K, 在B地区被调查750中有246人喜欢品牌K, 试估计两地区人们喜欢品牌K比例差的95%置信区间.

解 可以利用R中的内置函数`prop.test()`求两总体的比例差的置信区间, R中运行

```
> like<-c(478, 246)
> people<-c(1000, 750)
> prop.test(like, people)
```

得结果如下:

```
2-sample test for equality of proportions
with continuity correction
data:  like out of people
X-squared = 39.1394, df = 1, p-value = 3.946e-10
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1031446 0.1968554
sample estimates:
prop 1 prop 2
 0.478  0.328
```

可以看出A地区喜欢品牌K的人更多, 且A、B两地区喜欢品牌K的比例之差的95%的置信区间为(0.1031, 0.1969). ■

注:

- 同单样本一样, 上面的结果实际上是经过连续性修改后得到的;

- 由上面的公式, 我们也可以自己编写没有修正的两比例之间的区间估计函数`ratio.ci()`:

—— `ratio.ci()` 的定义 ——

```
ratio.ci<-function(x, y, n1, n2, conf.level=0.95){
  xbar1=x/n1;xbar2=y/n2
  xbar=xbar1-xbar2
  alpha = 1 - conf.level
  zstar=qnorm(1-alpha/2)
  *(xbar1*(1-xbar1)/n1+xbar2*(1-xbar2)/n2)^(1/2)
  xbar +c(-zstar, +zstar)
}
```

用到上例中, 运行

```
> ratio.ci(478,246,1000,750,conf.level=0.95)
[1] 0.1043112 0.1956888
```

因此, 这时两比例之差的95%的置信区间为(0.1043, 0.1957), 其长度修正下的结果略小了些.

§5.6 样本容量的确定

确定样本容量 n 是抽样中的一个重要问题. 样本容量抽取过少会丢失样本信息, 会导致误差太大而不满足要求; 若样本抽取太多, 虽然各种信息都包含了, 误差也降低了, 但同时会增加所需的人力、物力和费用开销. 所以权衡两者, 我们要抽取适当数量的样本.

5.6.1 估计正态总体均值时样本容量的确定

设总体 X 的均值为 μ , 方差为 σ^2 , 一般估计总体的均值时, 我们提出这样的精度要求, 以置信度 $1 - \alpha$, 允许均值的最大绝对误差为 d , 即

$$P(|\bar{X} - \mu| \leq d) = 1 - \alpha.$$

下面考虑总体 X 为正态(或近似正态)分布场合, 估计均值 μ 时所需的样本容量, 我们分两种情况进行讨论.

1. 总体方差 σ^2 已知

令 $\sigma^2 = \sigma_0^2$, 则由

$$\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \sim N(0, 1)$$

得

$$P\left(\frac{|\bar{X} - \mu|}{\sigma_0/\sqrt{n}} < \frac{d}{\sigma_0/\sqrt{n}}\right) = 1 - \alpha$$

所以

$$n = \left(\frac{z_{1-\frac{\alpha}{2}}\sigma_0}{d}\right)^2. \quad (5-6.1)$$

在R中可以定义如下的函数size.norm1()求样本容量:

```
size.norm1( )的定义
size.norm1<-function(d,var,conf.level) {
  alpha = 1 - conf.level
  ((qnorm(1-alpha/2)*var^(1/2))/d)^2
}
```

例 5.6.1 某地区有10000户, 拟抽取一个简单的样本调查一个月的平均开支, 要求置信度为95%, 最大允许误差为2, 根据经验, 家庭间开支的方差为500, 应抽取多少户进行调查?

```
> size.norm1(2,500,conf.level=0.95)
[1] 480.1824
```

所以应该抽取481户.

2. 总体方差 σ^2 未知

当 σ^2 未知时, 由

$$P\left(\frac{|\bar{X} - \mu|}{s/\sqrt{n}} < \frac{d}{s/\sqrt{n}}\right) = 1 - \alpha$$

得

$$n = \left(\frac{t_{1-\frac{\alpha}{2}}(n-1)s}{d}\right)^2. \quad (5-6.2)$$

注意到, $t_{1-\frac{\alpha}{2}}(n-1)$ 的值是随自由度 $(n-1)$ 而变化的, 也就是说 $t_{1-\frac{\alpha}{2}}(n-1)$ 的值原本就与样本容量 n 有关. 这样在 n 未确定之前 $t_{1-\frac{\alpha}{2}}(n-1)$ 的值也是未知的. 在这种情况下, 一般用尝试法, 先将一个非常大的自由度代入(相当于用 $z_{1-\alpha/2}$ 代替 $t_{1-\alpha/2}(n-1)$) 求出 n_1 , 然后再将 n_1 代入 $t_{1-\frac{\alpha}{2}}(n-1)$ 求出 n_2 , 重复此法直至先后两次所求得的 n 几乎相等为止, 最后的 n_2 就是要确定的样本容量.

在 **R** 中我们可以通过循环确定样本容量:

size.norm2() 的定义

```
size.norm2<-function(s,alpha,d,m){
  t0<-qt(alpha/2,m,lower.tail=FALSE)
  n0<-(t0*s/d)^2
  t1<-qt(alpha/2,n0,lower.tail=FALSE)
  n1<-(t1*s/d)^2
  while(abs(n1-n0)>0.5){
    n0<-(qt(alpha/2,n1,lower.tail=FALSE)*s/d)^2
    n1<-(qt(alpha/2,n0,lower.tail=FALSE)*s/d)^2
  }
  n1
}
```

说明: m 是事先给定的一个很大的数.

例 5.6.2 某公司生产了一批新产品, 产品总体服从正态分布, 现要估计这批产品的平均重量, 最大允许误差2, 样本标准差 $s = 10$, 试问 $\alpha = 0.01$ 下要抽取多少样本?

解 **R** 中的程序:

```
> size.norm2(10,0.01,2,100)
[1] 169.6658
```

也就是说在最大允许误差为2的时候应抽取170个样本. ■ 对估计量精度的要求还有别的提出方法, 比如要求均值的最大相对误差为 γ 或者是变异系数不超过 \sqrt{c} . 类似地, 我们可以求出样本容量的表达式, 据此通过 **R** 求解.

5.6.2 估计比例 p 时样本容量的确定

在样本容量较大的条件下, 样本比例 \hat{p} 的近似服从正态分布, 也即

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0, 1).$$

在置信水平 $1 - \alpha$ 下, 若允许比例的最大绝对误差为 d , 则由

$$P\left(\frac{|\hat{p} - p|}{\sqrt{p(1-p)/n}} < \frac{d}{\sqrt{p(1-p)/n}}\right) = 1 - \alpha,$$

从而

$$n = \left(\frac{z_{1-\frac{\alpha}{2}}}{d}\right)^2 p(1-p). \quad (5-6.3)$$

如果根据经验, 能给出 p 的一个粗略的估计值或者知道 p 的取值范围, 问题就能解决. (取值范围包括0.5时, 取 $p = 0.5$, 反之, 取接近0.5的值, 这样我们可以得到 n 的一个较为保守的值, 因为 $p(1-p) \leq 1/4$.) 如果对 p 没有任何先验知识时, 取 $p = 0.5$.

在R中我们这样实现:

size.bin()的定义

```
size.bin<-function(d, p, conf.level=0.95) {
  alpha = 1 - conf.level
  ((qnorm(1-alpha/2))/d)^2*p*(1-p)
}
```

例 5.6.3 某市一所重点大学历届毕业生就业率为90%, 试估计应届毕业生就业率, 要求估计误差不超过3%, 试问在 $\alpha = 0.05$ 下要抽取应届毕业生多少人?

解 R中的程序:

```
> size.bin(0.03, 0.9, 0.95)
[1] 384.1459
```

所以在 $\alpha = 0.05$ 下要抽取应届毕业生385人估计误差不超过3%. ■

第五章习题

5.1 设总体 X 是用无线电测距仪测量距离的误差, 它服从 (α, β) 上的均匀分布, 在200次测量中, 误差为 X_i 的次数有 n_i 次:

X_i	3	5	7	9	11	13	15	17	19	21
n_i	21	16	15	26	22	14	21	22	18	25

求 α, β 的矩法估计值(注: 这里的测量误差为 X_i 是指测量误差在 $(X_i - 1, X_i + 1)$ 间的代表值.)

5.2 为检验某自来水消毒设备的效果, 现从消毒后的水中随机抽取50L, 化验每升水中大肠杆菌的个数(假设1L水中大肠杆菌个数服从泊松分布), 其化验结果如下

大肠杆菌数/L	0	1	2	3	4	5	6
水的升数	17	20	10	2	1	0	0

试问平均每升水中大肠杆菌个数为多少时, 才能使上述情况的概率达到最大?

5.3 已知某种木材的横纹抗压力服从 $N(\mu, \sigma^2)$, 现对十个试件作横纹抗压力试验, 得数据如下(kg/cm^2):

482, 493, 457, 471, 510, 446, 435, 418, 394, 469

1) 求 μ 的置信水平为0.95的置信区间.

2) 求 σ 的置信水平为0.90的置信区间.

5.4 某卷烟厂生产两种卷烟A和B, 现分别对两种香烟的尼古丁含量进行6次试验, 结果如下

卷烟A	25	28	23	26	29	22
卷烟B	28	23	30	35	21	27

若香烟的尼古丁含量服从正态分布,

1) 问两种卷烟中尼古丁含量的方差是否相等?

2) 试求两种香烟的尼古丁平均含量差的95%置信区间.

5.5 比较两个小麦品种的产量, 选择22块条件相似地试验田, 采用相同的耕作方法做实验, 结果播种甲品种的12块实验田的单位面积产量和播种乙品种的12块实验田的单位面积产量分别为:

甲品种	628	583	510	554	612	523	530	615	573	603	334	564
乙品种	535	433	398	470	567	480	498	560	503	426	338	547

假定每个品种的单位面积产量均服从正态分布, 甲品种产量的方差为2140, 乙品种产量的方差为3250, 试求这两个品种平均面积产量差的置信水平为0.95的置信上限和置信水平为0.90的置信下限.

5.6 有两台机床生产同一型号的滚珠, 根据以往经验知, 这两台机床生产的滚珠直径都服从正态分布. 现分别从这两台机床生产的滚珠中随机地抽取7个和9个, 测得它们的直径如下(单位: mmm)

机床甲	15.2	14.5	15.5	14.8	15.1	15.6	14.7		
机床乙	15.2	15.0	14.8	15.2	15.0	14.9	15.1	14.8	15.3

试问机床乙生产的滚珠的方差是否比机床甲生产的滚珠直径的方差小?

5.7 某公司对本公司生产的两种自行车型号A、B的销售情况进行了了解, 随机选取了400人询问他们对A、B的选择, 其中有224人喜欢A, 试求顾客中喜欢A的人数比例 p 的置信水平为0.99的区间估计.

5.8 某公司生产了一批新产品, 产品总体服从正态分布, 现要估计这批产品的平均重量, 最大允许误差为1, 样本标准差 $s = 10$, 试问在0.95的置信度下至少要抽取多少个产品?

5.9 根据以往的经验, 船运大量玻璃器皿, 损坏率不超过5%. 现要估计某船中玻璃器皿的损坏率, 要求估计与真值间不超过1%, 且置信度为0.90, 那么要抽取多少样本验收可满足上述要求?

第六章 参数的假设检验

本章概要

- ◇ 假设检验的基本思想与检验的 p 值
- ◇ 正态总体均值和方差的假设检验
- ◇ 两正态总体均值和方差的比较
- ◇ 成对数据的假设检验
- ◇ 比例的检验与两比例的比较

上一章介绍了参数的点估计与区间估计的构造方法. 统计推断的另一重要内容是假设检验. 先对总体的某个未知参数或总体的分布形式作某种假设, 然后由抽取的样本提供的信息, 构造合适的统计量, 对所提供的假设进行检验, 以做出统计判断是接受假设还是拒绝假设, 这类统计推断问题称为假设检验问题, 前者称为参数假设检验, 后者称为非参数假设检验. 我们在本章和第七章中分别加以介绍.

§6.1 假设检验与检验的 p 值

6.1.1 假设检验的概念与步骤

统计假设

下面先通过几个例子来说明什么是假设检验.

例 6.1.1 微波炉在炉门关闭时的辐射量是一个重要的质量指标. 设该指标服从正态分布 $N(\mu, 0.1^2)$, 均值要求不超过0.12. 为检查近期产品的质量, 从某厂生产的微波炉中抽查了25台, 得其炉门关闭时辐射量的均值 $\bar{X} = 0.13$,

问该厂生产的微波炉炉门关闭时辐射量是否偏高?

本例是希望通过样本检验炉门关闭时辐射量是否高于0.12.

例 6.1.2 某车间用一台包装机包装精盐, 额定标准每袋净重500g, 设包装机包装出的盐每袋净重 $X \sim N(\mu, \sigma^2)$, 某天随机地抽取9袋, 称得净重为490, 506, 508, 502, 498, 511, 510, 515, 512. 问该包装机工作是否正常?

本例是希望通过样本检验包装机包装的盐的平均重量是否是500g.

以上两个例子都是参数的假设检验. 我们把施加于一个或多个总体的概率分布或参数上的假设称为统计假设, 简称假设. 所作的假设可以是真的, 也可能是假的. 为了判断一个统计假设是否正确, 需要作检验. 我们把判断统计假设是否正确的方法称为统计假设检验, 简称为统计检验.

假设检验的基本思想

1) 假设检验的基本思想

无论是怎样的假设, 假设检验的思想是一样的, 就是所谓概率性质的反证法. 其根据是实际推断原理: 小概率事件在一次实验中是几乎不可能发生的. 进一步讲, 要检验某假设 H_0 , 先假设 H_0 正确, 在此假设下构造某一事件A, 它在 H_0 为正确的条件下的概率很小, 例如 $P(A|H_0) = \alpha (= 0.05)$. 现在进行一次实验, 如果事件A发生了, 也就是说小概率事件在一次实验中居然发生了, 这与实际推断原理相矛盾, 这表明“假定 H_0 为正确”是错误的, 因而拒绝 H_0 ; 反之, 如果小概率事件没有发生, 我们就没有理由拒绝 H_0 , 通常就接受 H_0 .

通常称“结论”成立的假设为原假设(又称零假设), 记为 H_0 ; 与之对立的假设为备择假设(又称对立假设), 记为 H_1 . 我们将一个假设检验问题简记为 $H_0 \leftrightarrow H_1$. 例如, 例6.1.1中的假设检验问题为 $H_0: \mu \leq 0.12 \leftrightarrow H_1: \mu > 0.12$.

值得注意的是:

- 小概率事件在一次实验中发生与实际推断原理相矛盾, 这种矛盾并不是形式逻辑中的绝对矛盾, 因为“小概率事件在一次实验中几乎是不会发生的”, 并不意味着“小概率事件在一次实验中绝对不会发生”. 因此, 根据概率性质的反证法得出的接受 H_0 或拒绝 H_0 的决策, 并不等于我们证明了原假设 H_0 正确或错误, 而只是根据样本所提供的信息以一定的可靠程度认为 H_0 正确或错误.

- 原假设与备择假设并不对称或可以交换, 它们在假设检验中的地位是不同的. 原假设与备择假设的建立主要根据具体问题来决定的. 常把没有把握不能轻易肯定的命题作为备择假设, 而把没有充分理由不能轻易否定的命题作原假设, 只有理由充分时才拒绝它, 否则应予以保留.

2) 两类错误

从主观上讲, 我们总希望经过假设检验, 能作出正确的判断, 即若 H_0 确实为真, 则接受 H_0 ; 若 H_0 确实为假, 则拒绝 H_0 . 但在客观上, 我们是根据样本所确定的统计量之值来推断的, 由于样本的随机性, 在推断时就不免要犯错误. 因为当 H_0 正确时, 小概率事件也有可能发生而非绝对不可能发生, 这时我们却错误地否定了 H_0 . 这种“弃真”的错误, 称之为第一类错误; 由上所述, 犯第一类错误的概率为 $P(\text{拒绝 } H_0 | H_0 \text{ 为真}) = \alpha$. 还有可能犯“取伪”的错误, 称之为第二类错误, 就是当 H_0 不真, 但我们却接受了 H_0 . 犯第二类错误的概率为 $P(\text{接受 } H_0 | H_0 \text{ 为假}) = \beta$.

我们当然希望犯两类错误的概率都很小, 但是在样本容量固定时是办不到的. 通常把解决这一问题的原则简化成只对第一类错误的最大概率 α 加以限制, 而不考虑犯第二类错误的概率 β . 这种统计假设检验问题称为显著性检验, 并将犯第一类错误的最大概率 α 称为假设检验的显著性水平. 本书仅讨论显著性检验.

3) 检验步骤

先介绍接收域、拒绝域的概念: 对于一个检验问题 $H_0 \leftrightarrow H_1$, 当检验统计量 W 取某区域 C 中的值时, 我们拒绝原假设 H_0 , 则称区域 C 为 H_0 关于统计量 W 的拒绝域. 拒绝域的边界点称为临界点(或临界值). 当检验统计量 W 取某区域 C 中的值时, 我们无法拒绝(接受)原假设 H_0 , 则称区域 C 为 H_0 关于统计量 W 的接受域.

由以上的讨论, 我们归纳得到假设检验的主要步骤:

- 1) 提出原假设 H_0 与备择假设 H_1 ;
- 2) 选择检验统计量 W 并确定其分布;
- 3) 在给定的显著性水平下, 确定 H_0 关于统计量 W 的拒绝域;
- 4) 算出样本点对应的检验统计量的值;
- 5) 判断: 若统计量的值落在拒绝域内, 则拒绝 H_0 , 否则接受 H_0 .

6.1.2 检验的 p 值

定义 6.1.1 在一个假设检验问题中, 拒绝原假设 H_0 的最小显著性水平称为检验的 p 值.

从定义可知 p 值表示对原假设的怀疑程度, 或解释为首次拒绝原假设的概率. p 值越小, 表示原假设越可疑, 从而越应拒绝原假设. p 值的具体计算依赖于原假设、统计量的分布及其观测值. 现有的统计软件, 包括 \mathbf{R} 都提供了检验的 p 值.

引入检验的 p 值有明显的好处. 第一, 它比较客观地避免了事先确定显著性水平; 其次, 由检验的 p 值与人们心目中的显著性水平 α 进行比较可以很容易做出检验的结论: 如果 $\alpha \geq p$, 则在显著性水平 α 下拒绝 H_0 ; 如果 $\alpha < p$, 则在显著性水平 α 下保留 H_0 .

§6.2 单正态总体参数的检验

在实际中, 很多现象都可以近似地用正态分布描述, 因此关于正态分布参数均值和方差的检验, 是实际中常见的统计问题. 这一节先介绍单正态总体中的假设检验问题, 下一节考虑两正态总体中的假设检验问题.

假设总体 $X \sim N(\mu, \sigma^2)$, X_1, \dots, X_n 是来自此正态总体的一个样本, $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ 为其样本均值, $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ 为其样本方差.

6.2.1 均值 μ 的假设检验

1. 方差 σ^2 已知时 μ 的检验: Z 检验

设方差 $\sigma^2 = \sigma_0^2$ 已知, 考虑假设检验问题:

$$1) H_0: \mu = \mu_0 \longleftrightarrow H_1: \mu \neq \mu_0 \text{ (双边假设检验)}$$

$$2) H_0: \mu \leq \mu_0 \longleftrightarrow H_1: \mu > \mu_0 \text{ (单边假设检验)}$$

$$3) H_0: \mu \geq \mu_0 \longleftrightarrow H_1: \mu < \mu_0 \text{ (单边假设检验)}$$

在 $\mu = \mu_0$ 下可得

$$Z = \frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} \sim N(0, 1) \quad (6-2.1)$$

对于检验问题1), 若 \bar{X} 偏离 μ_0 (或左或右)均会倾向于拒绝原假设 H_0 , 从而接受对立假设 H_1 , 所以此问题的拒绝域为

$$C_1 = \{|Z| > z_{1-\alpha/2}\}.$$

对于检验问题2), 若 \bar{X} 大于 μ_0 , 则会倾向于拒绝原假设 H_0 , 从而接受对立假设 H_1 , 所以此问题的拒绝域为

$$C_2 = \{Z > z_{1-\alpha}\}.$$

对于检验问题3), 若 \bar{X} 大于 μ_0 , 则会倾向于拒绝原假设 H_0 , 从而接受对立假设 H_1 , 所以此问题的拒绝域为

$$C_3 = \{Z < -z_{1-\alpha}\}.$$

R程序在读入数据后, 还需要:

- 指定显著性水平 α 、原假设中的均值 μ_0 和已知的总体标准差 σ_0 ;
- 按上式计算出统计量 Z 的值;
- 计算 p 值.

设 Z_{obs} 表示统计量 Z 的观测值, 则对于上述三个假设检验问题, 相应的 p 值分别为:

$$1) P_1 = P(|Z| > |Z_{obs}|)$$

$$2) P_2 = P(Z > Z_{obs})$$

$$3) P_3 = P(Z < Z_{obs})$$

R中没有直接的函数来做方差已知时均值的检验, 需自己编写. 这里我们直接引用§5.2.1中做方差已知时均值的置信区间的函数`z.test()`.

例 6.2.1 在显著性水平 $\alpha = 0.05$ 下, 讨论例6.1.1的假设检验问题.

解 **R**程序如下:

```
> z.test(0.13, 25, 0.1, 0.05, u0=0.12, alternative="less")
```

运行结果为:

```
$mean
[1] 0.13
$z
[1] 0.5
$p.value
[1] 0.3085
$conf.int
[1] 0.0908 0.1692
```

结论: 因为 p 值=0.6915 $>$ $\alpha=0.05$, 故接收原假设, 认为炉门关闭时辐射量没有偏高. ■

2. 方差 σ^2 未知时 μ 的检验: t检验

设方差 σ^2 未知. 仍考虑假设检验问题1)、2)和3), 这时在 $\mu = \mu_0$ 下可得:

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(n-1) \quad (6-2.2)$$

由此得三个假设检验问题的拒绝域分别为:

$$C_1 = \{|T| > t_{1-\alpha/2}(n-1)\}$$

$$C_2 = \{T > t_{1-\alpha}(n-1)\}$$

$$C_3 = \{T < -t_{1-\alpha}(n-1)\}$$

与方差已知的情形相比, 我们并不需要复杂的编程, 直接利用R语言的`t.test()`函数就可完成原假设的检验. `t.test()`的调用格式见§5.2.1, 这里不再重复.

例 6.2.2 在显著性水平 $\alpha = 0.05$ 下, 讨论例6.1.2的假设检验问题.

解 R程序如下:

```
> salt<-c(490 , 506, 508, 502, 498, 511, 510, 515 , 512)
> t.test(salt, mu=500)
```

运行结果为:

```
One Sample t-test
data: salt
t = 2.198, df = 8, p-value = 0.05919
alternative hypothesis: true mean is not equal to 500
95 percent confidence interval:
 499.7 511.8
sample estimates:
mean of x
 505.8
```

结论: 因为 p 值 $=0.05919 > \alpha = 0.05$, 故接收原假设, 认为该包装机正常. ■

例 6.2.3 已知某种水样中 CaCO_3 的真值为 20.7mg/L , 现用某种方法重复测定该水样11次, CaCO_3 的含量为: 20.9, 20.41, 20.10, 20.00, 20.19, 22.60, 20.99, 20.41, 20, 23, 22. 问用该法测定的 CaCO_3 含量的均值与真值有无显著差异? (显著性水平为0.05)

解 R程序如下:

```
> CaCo3<-c(20.9, 20.41, 20.10, 20.00, 20.19,
           22.60, 20.99, 20.41, 20, 23, 22)
> t.test(CaCo3, mu=20.7)
```

运行结果为:

```
One Sample t-test
data: CaCo3
t = 0.8078, df = 10, p-value = 0.438
alternative hypothesis: true mean is not equal to 20.7
95 percent confidence interval:
 20.24 21.69
sample estimates:
mean of x
 20.96
```

结论: 因为 p 值 $=0.3125 > \alpha = 0.05$, 故认为此法所测定的水中 CaCO_3 的含量的均值与真值无显著差异, 故此法可信. ■

6.2.2 方差 σ^2 的检验: χ^2 检验

考虑假设检验问题:

$$1) H_0: \sigma^2 = \sigma_0^2 \longleftrightarrow H_1: \sigma^2 \neq \sigma_0^2 (\text{双边假设检验})$$

$$2) H_0: \sigma^2 \leq \sigma_0^2 \longleftrightarrow H_1: \sigma^2 > \sigma_0^2 (\text{单边假设检验})$$

$$3) H_0: \sigma^2 \geq \sigma_0^2 \longleftrightarrow H_1: \sigma^2 < \sigma_0^2 (\text{单边假设检验})$$

这时在 $\sigma^2 = \sigma_0^2$ 下可得:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1) \quad (6-2.3)$$

由此得三个假设检验问题的拒绝域分别为:

$$C_1 = \{\chi^2 \geq \chi_{1-\alpha/2}^2(n-1) \text{ 或 } \chi^2 \leq \chi_{\alpha/2}^2(n-1)\}$$

$$C_2 = \{\chi^2 \geq \chi_{1-\alpha}^2(n-1)\}$$

$$C_3 = \{\chi^2 \leq \chi_{\alpha}^2(n-1)\}$$

在R中没有直接的函数来做 χ^2 检验, 但§5.2.2中编写的函数`chisq.var.test()`可用于求单样本方差的检验.

例 6.2.4 检查一批保险丝, 抽出10根测量其通过强电流熔化所需的时间(单位: 秒)为: 42, 65, 75, 78, 59, 71, 57, 68, 54, 55. 假设熔化所需时间服从正态分布, 问能否认为熔化时间方差不超过80 (取 $\alpha = 0.05$).

解 R程序如下:

```
> time<-c(42, 65, 75, 78, 59, 71, 57, 68, 54, 55)
> chisq.var.test(time, 80, 0.05, alternative="less")
```

运行结果为:

\$var


```
[1] 121.8
$chi2
[1] 13.71
$p.value
[1] 0.8668
$conf.int
[1] 57.64 406.02
```

结论: 因为 p 值 $=0.8668 > \alpha = 0.05$, 故接收原假设, 认为熔化时间方差不超过80. ■

§6.3 两正态总体参数的检验

上节讨论了单个正态总体参数的显著性检验, 它是把样本统计量的观察值与原假设所提供的总体参数作比较, 这种检验要求我们事先能提出合理的参数假设值, 并对参数有某种意义的备择值, 但在实际工作中很难做到这一点, 因而限制了这种方法在实际中的应用. 实际中常常选择两个样本, 一个作为处理, 一个作为对照, 在两个样本之间作比较. 比如, 要比较某班男生的成绩是否比女生的高, 服用某种维生素的人是否比不服用的人不易感冒, 或判断它们之间是否存在明显显著的差异, 等等.

设总体 X 与 Y 独立, $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, X_1, \dots, X_{n_1} 是来自总体 X 的样本, $\bar{X} = \frac{1}{n_1} \sum_{i=1}^{n_1} X_i$ 为其样本均值, $S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$ 为其样本方差. Y_1, \dots, Y_{n_2} 是来自总体 Y 的样本, $\bar{Y} = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_i$ 为其样本均值, $S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$ 为其样本方差.

6.3.1 均值的比较: t 检验

设两正态总体的方差相等, 即 $\sigma_1^2 = \sigma_2^2 = \sigma^2$. 考虑假设检验问题:

- 1) $H_0 : \mu_1 = \mu_2 \longleftrightarrow H_1 : \mu_1 \neq \mu_2$ (双边假设检验)
- 2) $H_0 : \mu_1 \leq \mu_2 \longleftrightarrow H_1 : \mu_1 > \mu_2$ (单边假设检验)
- 3) $H_0 : \mu_1 \geq \mu_2 \longleftrightarrow H_1 : \mu_1 < \mu_2$ (单边假设检验)

这时在 $\mu_1 = \mu_2$ 下可得:

$$T = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})s^2}} \sim t(n_1 + n_2 - 2) \quad (6-3.1)$$

由此得三个假设检验问题的拒绝域分别为:

$$C_1 = \{|T| > t_{1-\alpha/2}(n_1 + n_2 - 2)\}$$

$$C_2 = \{T > t_{1-\alpha}(n_1 + n_2 - 2)\}$$

$$C_3 = \{T < -t_{1-\alpha}(n_1 + n_2 - 2)\}$$

在R语言中可以直接利用`t.test()`函数完成原假设的检验.

例 6.3.1 甲、乙两台机床分别加工某种轴承, 轴承的直径分别服从正态分布 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$, 从各自加工的轴承中分别抽取若干个轴承测其直径, 结果如表6.1所示. 设 $\sigma_1^2 = \sigma_2^2$, 问两台机床的加工精度有无显著差异?(取 $\alpha = 0.05$)

表 6.1 机床加工的轴的直径数据

总体	样本容量	直径
X(机床甲)	8	20.5 19.8 19.7 20.4 20.1 20.0 19.0 19.9
Y(机床乙)	7	20.7 19.8 19.5 20.8 20.4 19.6 20.2

解 R程序如下:

```
> x<-c(20.5, 19.8, 19.7, 20.4, 20.1, 20.0, 19.0, 19.9)
> y<-c(20.7, 19.8, 19.5, 20.8, 20.4, 19.6, 20.2)
> t.test(x, y, var.equal=TRUE)
```

运行结果为:

```
Two Sample t-test
data: x and y
t = -0.8548, df = 13, p-value = 0.4081
alternative hypothesis: true difference in means is not equal to 0
```

95 percent confidence interval:

-0.7684249 0.3327106

sample estimates:

mean of x mean of y

19.92500 20.14286

结论: 因为 p 值=0.4081 $>$ $\alpha = 0.05$, 故接收原假设, 认为两台机床的加工精度无显著差异. ■

6.3.2 方差的比较: F 检验

考虑假设检验问题:

1) $H_0: \sigma_1^2 = \sigma_2^2 \longleftrightarrow H_1: \sigma_1^2 \neq \sigma_2^2$ (双边假设检验)

2) $H_0: \sigma_1^2 \leq \sigma_2^2 \longleftrightarrow H_1: \sigma_1^2 > \sigma_2^2$ (单边假设检验)

3) $H_0: \sigma_1^2 \geq \sigma_2^2 \longleftrightarrow H_1: \sigma_1^2 < \sigma_2^2$ (单边假设检验)

这时在 $\sigma_1^2 = \sigma_2^2$ 下可得:

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1) \quad (6-3.2)$$

由此得三个假设检验问题的拒绝域分别为:

$$C_1 = \{F \geq F_{1-\alpha/2}(n_1 - 1, n_2 - 1) \text{ 或 } F \leq F_{\alpha/2}(n_1 - 1, n_2 - 1)\}$$

$$C_2 = \{F \geq F_{1-\alpha}(n_1 - 1, n_2 - 1)\}$$

$$C_3 = \{F \leq F_{\alpha}(n_1 - 1, n_2 - 1)\}$$

R语言中的`var.test()`函数可完成两样本的 F 检验. `var.test()`的调用格式见§5.3.2.

例 6.3.2 数据同例6.3.1, 问两台机床加工的轴的直径的方差是不是相同?

解 **R**程序如下:

```
> x<-c(20.5, 19.8, 19.7, 20.4, 20.1, 20.0, 19.0, 19.9)
> y<-c(20.7, 19.8, 19.5, 20.8, 20.4, 19.6, 20.2)
> var.test(x, y)
```

运行结果为:

```
F test to compare two variances
data:  x and y
F = 0.7932, num df = 7, denom df = 6, p-value = 0.7608
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1392675 4.0600387
sample estimates:
ratio of variances
 0.7931937
```

结论: 因为 p 值 $=0.7608 > \alpha = 0.05$, 故接收原假设, 认为两台机床加工的轴的直径的方差相同. ■

从本例也可知, 例6.3.1中做方差相同的假设是没有问题的. 以后在做两样本的均值检验时要先做方差齐性检验. 如果方差相等不满足, 则`t.test()`函数中使用选项`var.equal=FALSE`. 方差不等时均值检验问题还没有完全解决, 其近似检验方法请参看文献^[1].

§6.4 成对数据的t检验

上一节我们提过, 对一般情况下的两样本均值检验还没有完全解决. 本节考虑一种特殊的情况: 两样本成对数据的 t 检验. 所谓成对数据, 是指两个样本的样本容量相等, 且两个样本之间除均值之外没有另的差异. 例如比较某一班同一单元内容的第二次考试是否比第一次的高? 同一个人在服用某种维生素后是否比未服用之前不易感冒? 这就是成对数据的比较检验.

设 X_1, \dots, X_n 是来自总体 X 的样本, Y_1, \dots, Y_n 是来自总体 Y 的样本, 定义: $Z_i = X_i - Y_i (i = 1, 2, \dots, n)$, 记 $\mu = \mu_1 - \mu_2, \sigma^2 = \sigma_1^2 + \sigma_2^2$, 则 Z_1, Z_2, \dots, Z_n 为总体 $Z \sim N(\mu, \sigma^2)$ 的样本. 此时, μ_1 与 μ_2 的检验问题等价于 μ 的检验问题. 因此由单正态总体均值的假设检验知, 假设检验问题

$$1) H_0: \mu = \mu_0 \longleftrightarrow H_1: \mu \neq \mu_0 \text{ (双边假设检验)}$$

$$2) H_0: \mu \leq \mu_0 \longleftrightarrow H_1: \mu > \mu_0 \text{ (单边假设检验)}$$

3) $H_0: \mu \geq \mu_0 \longleftrightarrow H_1: \mu < \mu_0$ (单边假设检验)

的拒绝域分别为:

$$C_1 = \{|T| > t_{\alpha/2}(n-1)\}$$

$$C_2 = \{T > t_{\alpha}(n-1)\}$$

$$C_3 = \{T < -t_{\alpha}(n-1)\}$$

其中 $\mu = \mu_0$ 下

$$T = \frac{\bar{Z} - \mu_0}{S/\sqrt{n}} \sim t(n-1) \quad (6-4.1)$$

\bar{Z} 和 S 分别表示总体 Z 的样本均值和样本标准差.

在R语言中可以直接利用 `t.test()` 函数增加选项 `paired=TRUE` 完成原假设的显著性检验. 下面通过例子来说明具体的用法.

例 6.4.1 在针织品漂白工艺过程中, 要考虑温度对针织品断裂强力(主要质量指标)的影响. 为了比较70℃与80℃的影响有无差别, 在这两个温度下, 分别重复做了8次试验, 得数据如表6.2所示(单位: N): 根据经验, 温度对针织

表 6.2 温度对针织品断裂强力的影响数据

70℃时的强力	20.5 18.8 19.8 20.9 21.5 19.5 21.0 21.2
80℃时的强力	17.7 20.3 20.0 18.8 19.0 20.1 20.0 19.1

品断裂强度的波动没有影响. 问在70℃时的平均断裂强力与80℃时的平均断裂强力间是否有显著差别? 假定断裂强力服从正态分布($\alpha = 0.05$)

解 R程序如下:

```
> x<-c(20.5, 18.8, 19.8, 20.9, 21.5, 19.5, 21.0, 21.2)
> y<-c(17.7, 20.3, 20.0, 18.8, 19.0, 20.1, 20.0, 19.1)
> t.test(x, y, paired=TRUE)
```

运行结果为:

```
Paired t-test
data: x and y
```

```
t = 1.8002, df = 7, p-value = 0.1149
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3213757  2.3713757
sample estimates:
mean of the differences
          1.025
```

结论: 因为 p 值 $=0.1149 > \alpha = 0.05$, 故接收原假设, 认为在 70°C 时的平均断裂强力与 80°C 时的平均断裂强力间无显著差别。 ■

除了用`t.test()`函数完成原假设的检验外, **R**中还可以用**DAAG**包中的`onesamp()`函数来完成检验, `onesamp()`函数的调用格式如下:

—— `onesamp()` 的调用格式 ——

```
onesamp(dset=corn, x="unsprayed", y="sprayed", xlab=NULL,
        ylab=NULL, dubious=NULL, conv=NULL, dig=2)
```

说明: `dset`为有两列的数据框或矩阵; `x`为处于“predictor”地位的列名; `y`为处于“response”地位的列名。

下面用`onesamp()`函数来做上面的例子。

R程序如下:

```
> data.x<-c(20.5, 18.8, 19.8, 20.9, 21.5, 19.5, 21.0, 21.2)
> data.y<-c(17.7, 20.3, 20.0, 18.8, 19.0, 20.1, 20.0, 19.1)
> z<-data.frame(data.x, data.y)
> onesamp(z, x="data.y", y="data.x")
```

运行结果为:

```
data.x 0.941124 0.8876132 1.610457
      One Sample t-test
data:  d
t = 1.8002, df = 7, p-value = 0.1149
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
```

```

-0.3213757  2.3713757
sample estimates:
mean of x
      1.025

```

所得结论与前面相同.

§6.5 单样本比率的检验

设 X_1, X_2, \dots, X_n 为来自二点分布(贝努里分布) $\text{binom}(1, p)$ 的样本, 则 $T = \sum_{i=1}^n X_i \sim \text{binom}(n, p)$.

6.5.1 比率 p 的精确检验

考虑假设检验问题:

- 1) $H_0 : p = p_0 \longleftrightarrow H_1 : p = p_0$ (双边假设检验)
- 2) $H_0 : p \leq p_0 \longleftrightarrow H_1 : p > p_0$ (单边假设检验)
- 3) $H_0 : p \geq p_0 \longleftrightarrow H_1 : p < p_0$ (单边假设检验)

基于统计量 $T = \sum_{i=1}^n X_i$ 作检验, 上述三个检验问题拒绝域分别有如下形式:

$$C_1 = \{T \leq c_1 \text{ 或 } T \geq c_2\}, \quad c_1 < c_2;$$

$$C_2 = T \geq c;$$

$$C_3 = T \leq c'.$$

为获得水平为 α 的检验, 需要定出各拒绝域中的临界值 c, c', c_1, c_2 . 下面仅以检验问题2)来说明两种确定临界值的方法.

利用二项分布来确定临界值

对于检验问题2), c 是满足下式的最小整数:

$$P(T \geq c) = \sum_{i=c}^n \binom{n}{i} p_0^i (1-p_0)^{n-i} \leq \alpha \quad (6-5.1)$$

用 F 分布来确定临界值

根据二项分布与 F 分布之间的关系

$$\sum_{i=c}^n \binom{n}{d} p_0^i (1-p_0)^{n-i} = F\left(\frac{n_2}{n_1} \frac{p_0}{1-p_0}; n_1, n_2\right) \quad (6-5.2)$$

右端是自由度为 n_1, n_2 的 F 分布的分布函数在 $\frac{n_2}{n_1} \frac{p_0}{1-p_0}$ 处的值, $n_1 = 2c, n_2 = 2(n-c+1)$. 这样为求出使(6-5.1)式成立的最小整数 c 等价于求使 $F_\alpha(n_1, n_2) \geq \frac{n_2 p_0}{n_1(1-p_0)}$ 成立的最小整数 c .

R语言中的`binom.test()`函数可完成原假设的检验. `binom.test()`的调用格式见§5.4.

6.5.2 比率 p 的近似检验

在样本容量较大时, 比例 p 的抽样分布可近似地服从正态分布, 因此我们可将问题转化为正态分布处理. 考虑上述假设检验问题, 在 $p = p_0$ 条件下构造统计量

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim N(0, 1) \quad (6-5.3)$$

其中 $\hat{p} = \frac{T}{n}$. 由此上述三个检验问题的拒绝域分别为:

- 1) $C_1 = \{|Z| > z_{1-\frac{\alpha}{2}}\}$
- 2) $C_2 = \{Z > z_{1-\alpha}\}$
- 3) $C_3 = \{Z < -z_{1-\alpha}\}$

R语言中的`prop.test()`函数可完成原假设的检验. `prop.test()`的调用格式见§5.4.

例 6.5.1 某产品的优质品率一直保持在40%, 近期技监部门抽查了12件产品, 其中优质品为5件, 问在 $\alpha = 0.05$ 水平上能否认为其优质品率仍保持在40%?

解 由于本例的样本容量不大, 不适合用大样本的方法来处理, 故我们对 p 做精确检验. **R**程序如下:


```
> binom.test(c(7, 5), p=0.4)
```

运行结果为:

```
Exact binomial test
data:  c(7, 5)
number of successes = 7, number of trials = 12, p-value = 0.2417
alternative hypothesis:
true probability of success is not equal to 0.4
95 percent confidence interval:
 0.2766697 0.8483478
sample estimates:
probability of success
      0.5833333
```

结论: 因为 p 值 $=0.2417 > \alpha = 0.05$, 故接收原假设, 认为该产品的优质品率仍保持在40%.

同样的, 我们也可以用`prop.test()`进行检验. **R**程序如下:

```
> prop.test(7, 12, p=0.4, correct=TRUE)
```

运行结果为:

```
1-sample proportions test with continuity correction
data:  7 out of 12, null probability 0.4
X-squared = 1.0035, df = 1, p-value = 0.3165
alternative hypothesis: true p is not equal to 0.4
95 percent confidence interval:
 0.2859928 0.8350075
sample estimates:
      p
0.5833333
Warning message:
In prop.test(7, 12, p = 0.4, correct = TRUE) :
  Chi-squared approximation may be incorrect
```

结论: 因为 p 值 $=0.3165 > \alpha = 0.05$, 故接收原假设, 认为该产品的优质品率仍保持在40%.

说明: 当样本容量较小而做近似检验时, **R**输出的结果会有警告信息(warning message):Chi-squared近似算法有可能不准. 在**R**中, 当样本容量大于20时不会出现这样的警告信息. 通常, 我们一般在样本容量大于30时做大样本近似. ■

例 6.5.2 某大学随机调查120名男同学, 发现有35人喜欢看武侠小说, 问可否认为该大学有四分之一的男同学喜欢看武侠小说?(取 $\alpha = 0.05$)

解 **R**程序如下:

```
> prop.test(35, 120, p=0.25, conf.level=0.975, correct=TRUE)
```

运行结果为:

```
1-sample proportions test with continuity correction
data: 35 out of 120, null probability 0.25
X-squared = 0.9, df = 1, p-value = 0.3428
alternative hypothesis: true p is not equal to 0.25
97.5 percent confidence interval:
 0.2049114 0.3958969
sample estimates:
      p
0.2916667
```

结论: 因为 p 值 $=0.3428 > \alpha = 0.05$, 故接收原假设, 认为该大学有四分之一的男同学喜欢看武侠小说. ■

§6.6 两样本比率的检验

设有两总体 X 与 Y 相互独立(总体容量都较大), 从中分别抽取 n_1 和 n_2 个(n_1, n_2 也较大)观察, 结果发现其中各有 x_1 和 x_2 个具有某种性质. 设总体 X 与 Y 中具有上述待性的比率分别为 p_1 和 p_2 , 我们的目的是要估计对下面的假设作出检验.

1) $H_0 : p_1 = p_2 \longleftrightarrow H_1 : p_1 \neq p_2$ (双边假设检验)

2) $H_0 : p_1 \leq p_2 \longleftrightarrow H_1 : p_1 > p_2$ (单边假设检验)

3) $H_0 : p_1 \geq p_2 \longleftrightarrow H_1 : p_1 < p_2$ (单边假设检验)

两个总体比例 p_1 和 p_2 的极大似然估计分别为 $\hat{p}_1 = \frac{x_1}{n_1}, \hat{p}_2 = \frac{x_2}{n_2}$. 由 §5.5, 若 n_1 和 n_2 较大, 则 \hat{p}_1, \hat{p}_2 近似地服从正态分布:

$$\hat{p}_1 \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right), \quad \hat{p}_2 \sim N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right).$$

在 $p_1 = p_2$ 下, 有

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{(n_1+n_2)\hat{p}(1-\hat{p})}{n_1 n_2}}} \sim N(0, 1), \quad (6-6.1)$$

其中 $\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$. 由此可知, 上述三个检验问题的拒绝域分别为:

1) $C_1 = \{|Z| > z_{1-\frac{\alpha}{2}}\}$

2) $C_2 = \{Z > z_{1-\alpha}\}$

3) $C_3 = \{Z < -z_{1-\alpha}\}$

R语言中的 `prop.test()` 函数可完成原假设的检验.

例 6.6.1 某高校随机抽取了102个男学生与135个女学生调查家中有无计算机, 调查结果为23个男学生与25个女学生家中拥有计算机. 问在 $\alpha = 0.05$ 水平上, 能否认为男、女学生家中拥有计算机的比率一致?

解 **R**程序如下:

```
> success<-c(23, 25)
> total<-c(102, 135)
> prop.test(success, total)
```

运行结果为:

2-sample test for equality of proportions

```
with continuity correction
data:  success out of total
X-squared = 0.3615, df = 1, p-value = 0.5477
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.07256476  0.15317478
sample estimates:
   prop 1    prop 2 
0.2254902 0.1851852
```

结论: 因为 p 值 $=0.5477 > \alpha = 0.05$, 故接收原假设, 认为该大学的男、女学生家中拥有计算机的比率一致. ■

第六章习题

6.1 有一批枪弹, 出厂时, 其初速 $\nu \sim N(950, \sigma^2)$ (单位: m/s). 经过较长时间储存, 取9发进行测试, 得样本值 (单位: m/s) 如下: 914, 920, 910, 934, 953, 940, 912, 924, 930. 据经验, 枪弹储存后其初速仍服从正态分布, 且标准差不变, 问是否可认为这批枪弹的初速有显著降低? ($\alpha = 0.01$)

6.2 已知维尼纶纤度在正常条件下服从正态分布, 且标准差为0.048. 从某天生产的产品中抽取5根纤维, 测得其纤度为: 1.32, 1.55, 1.36, 1.40, 1.1, 问这天抽取的维尼纶纤度的总体标准差是否正常? ($\alpha = 0.05$)

6.3 下面给出两种型号的计算器充电以后所能使用的时间(单位:小时)的观测值

型号A	5.5	5.6	6.3	4.6	5.3	5.0	6.2	5.8	5.1	5.2	5.9	
型号B	3.8	4.3	4.2	4.9	4.5	5.2	4.8	4.5	3.9	3.7	3.6	2.9.

设两样本独立且数据所属的两个总体的密度函数至多差一个平移量. 试问能否认为型号A的计算器平均使用时间比型号B来得长? ($\alpha = 0.01$)

6.4 测得两批电子器件的样本的电阻(Ω)为

A批(x)	0.140	0.138	0.143	0.142	0.144	0.137
B批(y)	0.135	0.140	0.142	0.136	0.138	0.130

设这两批器材的电阻值分别服从正态分布 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$, 且两样本独立,

(1) 试检验两个总体的方差是否相等? ($\alpha = 0.01$)

(2) 试检验两个总体的均值是否相等? ($\alpha = 0.05$)

6.5 有人称某地成年人中大学毕业生比例不低于30%, 为检验之, 随机调查该地15名成年人, 发现有3名大学毕业生, 取 $\alpha = 0.05$, 问该人的看法是否成立?

第七章 非参数的假设检验

本章概要

- ◇ 单一样本的检验
- ◇ 两样本比较与检验
- ◇ 多样本的比较与检验

上章讲的参数假设检验是在假设总体分布已知的情况下进行的. 但在实际生活中, 那种对总体的分布的假定并不是能随便作出的. 数据并不是来自所假定分布的总体, 或者, 数据根本不是来自一个总体; 还有可能数据因为种种原因被严重污染. 这样, 在假定总体分布已知的情况下进行推断的做法就可能产生错误甚至灾难性的结论. 于是, 人们希望在不对总体分布作出假定的情况下, 尽量从数据本身来获得所需要的信息, 这就是非参数统计推断的宗旨. 本章分别就单一样本、两样本及多样本的位置参数与尺度参数给出一些非参数的检验方法.

§7.1 单总体位置参数的检验

设 X_1, X_2, \dots, X_n 为来自总体 X 的容量为 n 的样本, 在有了样本观测值 x_1, x_2, \dots, x_n 之后, 很自然地想要知道它所代表的总体的“中心”在哪里? 它所代表的总体的分布是否与我们所希望的分布一样? 这些问题中不涉及分布具体形式的假定, 因此属于非参数的假设检验问题. 我们先考虑前一问题, 分别介绍两常用的中位数符号检验和对称中心的Wilcoxon符号秩检验, 后面一节再介绍分布的拟合优度检验.

7.1.1 中位数的符号检验

我们知道在总体为正态分布时, 要检验其均值是否为 μ , 是用 t 检验. 它的检验统计量 $T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ 在零假设成立时服从自由度为 $n - 1$ 的 t 分布. 但是, t 检验并不稳健, 在不知道总体分布时, 特别是在小样本场合, 运用 t 检验就可能有风险. 这时就要考虑使用非参数方法对分布的中心进行检验, 如本小节讨论的中位数的符号检验.

这一小节使用总体 X 的中位数 M 作为分布中心, 即 M 满足: $P(X < M) = P(X > M) = \frac{1}{2}$.

考虑假设检验问题:

$$1) H_0 : M = M_0 \longleftrightarrow H_1 : M > M_0 (\text{单边假设检验})$$

$$2) H_0 : M = M_0 \longleftrightarrow H_1 : M < M_0 (\text{单边假设检验})$$

$$3) H_0 : M = M_0 \longleftrightarrow H_1 : M \neq M_0 (\text{双边假设检验})$$

符号检验的检验统计量为:

$$S^+ = \# \{X_i : X_i - M_0 > 0, i = 1, 2, \dots, n\} \quad (7-1.1)$$

其中 $\#$ 表示计数, 即 S^+ 是集合 G 中的元素的个数, 其中 G 是使得 $X_i - M > 0$ 成立的 $X_i (i = 1, 2, \dots, n)$ 构成的集合. S^+ 也可以等价地表示为:

$$S^+ = \sum_{i=1}^n u_i, \quad u_i = \begin{cases} 1, & X_i - M_0 > 0, \\ 0, & \text{其它.} \end{cases}, i = 1, 2, \dots, n \quad (7-1.2)$$

由上面的假设可知:

$$S^+ \sim b(n, \frac{1}{2}).$$

由此上述三个假设检验问题的拒绝域分别为:

$$C_1 = \{S^+ \geq C\}, \text{ 其中 } C = \inf\{C^* : (\frac{1}{2})^n \sum_{i=C}^n \binom{n}{i} \leq \alpha\}$$

$$C_2 = \{S^+ \leq D\}, \text{ 其中 } D = \sup\{D^* : (\frac{1}{2})^n \sum_{i=0}^D \binom{n}{i} \leq \alpha\}$$

$C_3 = \{S^+ \geq C \text{ 或 } S^+ \leq D\}$, 其中 C, D 满足:

$$C = \inf \left\{ C^* : \left(\frac{1}{2} \right)^n \sum_{i=C}^n \binom{n}{i} \leq \frac{\alpha}{2} \right\}, D = n - C \quad (7-1.3)$$

注: 在实际问题中可能有某一些观察值 x_i 正好等于 M_0 , 一般采用的方法是把这些正好等于 M_0 的观察值舍去, 并相应地减少样本容量的 n 值.

另外, 因为 $E(S^+) = \frac{n}{2}$, $\text{Var}(S^+) = \frac{n}{4}$, 所以当 n 比较大时, 有

$$Z = \frac{S^+ - \frac{n}{2}}{\sqrt{n/2}} \sim N(0, 1). \quad (7-1.4)$$

因为正态分布是连续性的, 所以在离散的二项分布近似中, 要用连续性修正量, 即用

$$Z' = \frac{S^+ - \frac{n}{2} \pm 0.5}{\sqrt{n/2}} \sim N(0, 1). \quad (7-1.5)$$

这里分子的 \pm 处, 当 $S^+ < \frac{n}{2}$ 时取加号, 当 $S^+ > \frac{n}{2}$ 时取减号.

在 **R** 中没有直接的函数来做符号检验, 需要编写函数来做检验. 借助函数 `binom.test` 函数(见§5.4) `sign.test()` 定义如下:

—— `sign.test()` 的定义 ——

```
sign.test<-function(x, m0, alpha=0.05, alter="two.sided"){
  p<-list( )
  n<-length(x)
  sign<-as.numeric(x>=m0)
  s<-sum(sign)
  result<-binom.test(s, n, p=0.5, alternative=alter,
                     conf.level=alpha)
  p$p.value=result$p.value
  p
}
```

说明: `alter` 的取值为 “two.sided” 或 “greater”, “two.sided” 表示双边检验, “greater” 表示单边检验.

例 7.1.1 在某保险种类中, 一次关于2006年的索赔数额(单位:元)的随机抽样为(按升幂排列):

4632, 4728, 5052, 5064, 5484, 6972, 7696, 9048,

14760, 15013, 18730, 21240, 22836, 52788, 67200.

已知2005年的索赔数额的中位数为6064元. 问2006年索赔的中位数与前一年是否有所变化? ($\alpha = 0.05$)

解 R程序如下:

```
> insure<-c(4632, 4728, 5052, 5064, 5484, 6972, 7696, 9048,
            14760, 15013, 18730, 21240, 22836, 52788, 67200)
> sign.test(insure,6064)
```

运行结果为:

```
$p.value
[1] 0.3017578
```

结论: 因为 p 值=0.3017578 $>$ $\alpha = 0.05$, 故接收原假设, 认为2006年索赔的中位数与前一年没有发生变化. ■

7.1.2 Wilcoxon符号秩检验

符号检验利用了观察值和原假设的中心位置之差的符号来进行检验, 但是它并没有利用这些差的大小(体现于差的绝对值的大小)所包含的信息. 不同的符号代表了在中心位置的哪一边, 而差的绝对值的秩的大小代表了距离中心的远近. Wilcoxon符号秩检验把这两者结合起来, 所以要比仅仅利用符号的符号检验要更有效.

Wilcoxon符号秩检验使用总体 X 的对称中心 M 作为分布中心, 即总体 X 的分布 $F(x)$ 关于 M 对称, M 满足: $F(M - x) = 1 - F(x - M), \forall x \in R$. 在此我们还要求 X 是连续型的.

仍考虑上一小节的假设检验问题. Wilcoxon符号秩检验的检验统计量为:

$$W^+ = \sum_{i=1}^n u_i R_i \quad (7-1.6)$$

其中 u_i 的定义同(7-1.2)式, R_i 为 $|X_i|$ 在样本绝对值 $|X_1|, |X_2|, \dots, |X_n|$ 中的秩.

由此以上三个假设检验问题的拒绝域分别为:

$C_1 = \{W^+ \geq C\}$, 其中 C 满足: $C = \inf\{C^* : P(W^+ \geq C^*) \leq \alpha\}$

$C_2 = \{W^+ \leq D\}$, 其中 D 满足: $D = \sup\{D^* : P(W^+ \leq D^*) \leq \alpha\}$

$C_3 = \{W^+ \geq C \text{ 或 } S^+ \leq D\}$, 其中 C, D 满足:

$$C = \inf\{C^* : P(W^+ \geq C^*) \leq \frac{\alpha}{2}\}, D = \sup\{D^* : P(W^+ \leq D^*) \leq \frac{\alpha}{2}\}$$

为求上述检验 p 值, 需要知道 W^+ 的分布. 我们有

定理 7.1.1 令 $S = \sum_{i=1}^n iu_i$, 则在总体的分布关于原点 0 对称时, W^+ 与 S 同分布.

定理 7.1.2 在总体的分布关于原点 0 对称时, W^+ 的概率分布为:

$$P(W^+ = d) = P\left(\sum_{i=1}^n u_i R_i = d\right) = \frac{t_n(d)}{2^n},$$

$$d = 1, 2, \dots, \frac{n(n+1)}{2} \quad (7-1.7)$$

其中 $t_n(d)$ 表示从 $1, 2, \dots, n$ 这 n 个数中任取若干个其和恰为 d 的取法总数.

定理 7.1.3 在总体的分布关于原点 0 对称时, W^+ 服从对称分布, 对称中心为 $0, 1, 2, \dots, \frac{n(n+1)}{2}$ 的中点 $\frac{n(n+1)}{4}$.

有了以上三个定理, 我们就可以计算 p 值了(从略). 另外, 由于

$$E(W^+) = \frac{n(n+1)}{4}, \text{Var}(W^+) = \frac{n(n+1)(2n+1)}{24},$$

故当 n 比较大时, 有

$$Z = \frac{W^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim N(0, 1). \quad (7-1.8)$$

R 中的函数 `wilcoxon.test()` 可完成原假设的检验, 调用格式如下:

wilcoxon.test() 的调用格式

```
wilcox.test(x, y=NULL, alternative=c("two.sided", "less", "greater"),
            mu=0, paired = FALSE, exact = NULL, correct = TRUE,
```

```
conf.int = FALSE, conf.level = 0.95, ...)
```

说明: `exact` 表示是否算出准确的 p 值; `correct` 表示大样本时是否做连续性修正.

例 7.1.2 用 Wilcoxon 检验对例 7.1.1 的数进行检验.

解 R 程序如下:

```
> insure<-c(4632, 4728, 5052, 5064, 5484, 6972, 7696, 9048,
            14760, 15013, 18730, 21240, 22836, 52788, 67200)
> wilcox.test(insure,mu=6064,conf.int = TRUE)
```

运行结果为:

```
Wilcoxon signed rank test
data:  insure V = 101, p-value = 0.01807 alternative hypothesis:
true mu is not equal to 6064 95 percent confidence interval:
 6840 28926
sample estimates: (pseudo)median
 13065
```

结论: 因为 p 值 $= 0.01807 < \alpha = 0.05$, 故拒绝原假设, 认为 2006 年索赔的中位数与前一年有变化. 根据 95% 的置信区间, 2006 年索赔的中位数有所增加; 并且给出了一个 (伪) 中位数 13065. 这与中位数的符号检验所得的结果不同, 说明了 Wilcoxon 符号秩检验比符号检验利用了更多的信息, 检验应更有效. ■

§7.2 分布的一致性检验: χ^2 检验

在给定一些数据之后, 我们往往会假设它们来自某种分布, 但是这种假设对不对呢? 这一节我们讨论这一问题.

考虑假设检验问题

$$H_0 : F(x) = F_0(x) \longleftrightarrow H_1 : F(x) \neq F_0(x)$$

在随机变量 X 的取值范围 $[a, b]$ (a 可为 $-\infty$, b 可为 ∞) 内选取 $m-1$ 个实数 $a =$

$a_0 < a_1 < a_2 < \cdots < a_{m-1} < a_m = b$, 它们将 $[a, b]$ 分为 m 个小区间 $A_i = [a_{i-1}, a_i)$, 记 $p_{i0} = F_0(a_i) - F_0(a_{i-1})$

设 (x_1, x_2, \cdots, x_n) 为来自总体 $F(x)$ 的容量为 n 的一组样本观测值, n_i 为观测值落入 A_i 的频数, $\sum_{i=1}^m n_i = n$. 若 H_0 成立, 则实际频数 n_i 与理论频数 np_{i0} 比较接近, 因此分布的拟合优度检验可转化为分类数据的实际频数与理论频数的一致性检验. 下面的定理为此提供了理论依据.

定理 7.2.1 (Pearson定理)

1) 若 $F_0(x)$ 完全已知(不带有未知参数), 则当 H_0 成立时, 统计量

$$\chi^2 = \sum_{i=1}^m \frac{(n_i - nP_{i0})^2}{nP_{i0}} \sim \chi^2(m-1).$$

2) 若 $F_0(x) = F_0(x, \theta_1, \theta_2, \cdots, \theta_r)$ 中含有 r 个未知参数 $\theta_1, \theta_2, \cdots, \theta_r$, 它们的极大似然估计为 $\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_r$. 令 $\hat{p}_{i0} = F_0(a_i, \hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_r) - F_0(a_{i-1}, \hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_r)$, $i = 1, 2, \cdots, m$, 则

$$\chi^2 = \sum_{i=1}^m \frac{(n_i - n\hat{P}_{i0})^2}{n\hat{P}_{i0}} \sim \chi^2(m-r-1),$$

其中 m 表示种类数, r 表示参数个数.

由定理7.2.1知, 上述检验问题的拒绝域为 $C = \{\chi^2 > \chi_{1-\alpha}^2(m-1)\}$.

R中函数`chisq.test()`可完成原假设的检验. `chisq.test()`的调用格式如下:

`chisq.test()`的调用格式

```
chisq.test(x, y = NULL, correct = TRUE, p = rep(1/length(x), length(x)),
           rescale.p = FALSE, simulate.p.value = FALSE, B = 2000)
```

说明: x 为向量或矩阵. 若 x 是一维的且 y 不给出($y = \text{NULL}$), 则`chisq.test()`函数用于本节分布的拟和优度检验, 这时是检验总体概率是否与给定的 p 相同, p 缺省表示进行等可能性检验; x 与 y 同时给出时则进行7.3.1小节介绍的列联表检验.

例 7.2.1 某箱子中盛有10种球, 现在从中有放回地随机抽取200个, 其中第 i 种球共取得 ν_i 个, 数据记录在表7.1. 问箱子中这10种球的比例是否一

样? $(\alpha = 0.05)$

表 7.1 10种球的数目

种别	ν_i	种别	ν_i	种别	ν_i
1	35	5	17	9	30
2	16	6	19	10	14
3	15	7	11		
4	17	8	16		

解 R程序如下:

```
> v<-c(35,16,15,17,17,19,11,16,30,24)
> chisq.test(v)
```

运行结果为:

```
Chi-squared test for given probabilities
data:  v
X-squared = 24.9, df = 9, p-value = 0.003084
```

结论: 因为 p 值= $0.003084 < \alpha = 0.05$, 故拒绝原假设, 认为箱子中的10种球的比例不一样. ■

例 7.2.2 卢瑟福和盖革作了一个著名的实验, 他们观察了长为7.5秒的时间间隔里由某块放射物质放出的到达某个计数器的 α 质点数, 共观察了2608次. 表7.2的第一列给出的是质点数 i , 第二列表示相应的频数 n_i . 试问这种分布规律是否服从泊松分布? $(\alpha = 0.05)$

解 在R中没有直接算带参数的拟合检验函数, 故要根据具体问题自己编程.

首先计算参数 λ 的极大似然估计

R程序如下:

```
> x<-c(0,1,2,3,4,5,6,7,8,9,10)
```

表 7.2 放射物质放出的 α 质点数与频数

质点数 i	频数 n_i	质点数 i	频数 n_i	质点数 i	频数 n_i
0	57	4	532	8	45
1	203	5	408	9	27
2	383	6	273	10	16
3	525	7	239		

```

> y<-c(57,203,383,525,532,408,273,139,45,27,16)
> options(digits=3)
> likely<-function(lambda=3){
  -sum(y*dpois(x, lambda=lambda, log=TRUE))
}
> mle(likely)

```

运行结果为:

```

Call:
mle(minuslogl = likely) Coefficients: lambda 3.87

```

由于函数`chisq.test()`无法调整因参数估计引起的自由度调整, 因此需要编程计算检验统计量及 p 值, **R**程序如下:

```

> chisq.fit<-function(x, y, r){
  options(digits=4)
  result<-list( )
  n<-sum(y)
  prob<-dpois(x,3.87,log=FALSE)
  y<-c(y,0)
  m<-length(y)
  prob<-c(prob,1-sum(prob))
  result$chisq<-sum((y-n*prob)^2/(n*prob))
  result$p.value<-pchisq(result$chisq,m-r-1,lower.tail=FALSE)
  result
}

```

```

    }
> x<-c(0,1,2,3,4,5,6,7,8,9,10)
> y<-c(57,203,383,525,532,408,273,139,45,27,16)
> chisq.fit(x,y,1)

```

运行结果为:

```

$chisq
[1] 20.55
$p.value
[1] 0.02442

```

结论: 因为 p 值=0.02442 $< \alpha = 0.05$, 故拒绝原假设, 认为该分布规律不服从泊松分布. ■

§7.3 两总体的比较与检验

在单样本问题中, 人们想要检验的是总体的中心是否等于一个已知的值. 但在实际问题中, 更受注意的往往是比较两个总体的位置参数; 比如, 两种训练方法中哪一种更出成绩, 两种汽油中哪一种污染更少, 两种市场营销策略中哪种更有效等等.

7.3.1 χ^2 独立性检验

若随机变量 X, Y 的分布函数分别为 $F_1(x)$ 和 $F_2(y)$, 且联合分布为 $F(x, y)$, 则 X 与 Y 的独立性归结为假设检验问题:

$$H_0 : F(x, y) = F_1(x)F_2(y) \longleftrightarrow H_1 : F(x, y) \neq F_1(x)F_2(y).$$

若 X 与 Y 为分类变量, 其中 X 的取值为 X_1, X_2, \dots, X_r , Y 的取值为 Y_1, Y_2, \dots, Y_s , 将 X 与 Y 的各种情况的组合用一张 $r \times s$ 列联表表示, 称为 $r \times s$ 二维列联表, 如表7.3所示, 表中 n_{ij} 表示 n 个随机试验的结果中 X 取 X_i 及 Y 取 Y_j 的频数, $\sum_{i=1}^r \sum_{j=1}^s n_{ij} = n$,

$$n_{i.} = \sum_{j=1}^s n_{ij}, i = 1, 2, \dots, r, \text{表示各行之和}$$

表 7.3 $r \times s$ 列联表

	Y_1	Y_2	\cdots	Y_s	总和
X_1	n_{11}	n_{12}	\cdots	n_{1s}	$n_{1\cdot}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
X_r	n_{r1}	n_{r2}	\cdots	n_{rs}	$n_{r\cdot}$
总和	$n_{\cdot 1}$	$n_{\cdot 2}$	\cdots	$n_{\cdot s}$	n

$$n_{\cdot j} = \sum_{i=1}^r n_{ij}, j = 1, 2, \dots, s, \text{表示各列之和}$$

令 $p_{ij} = P(X = X_i, Y = Y_j)$, $p_{i\cdot} = P(X = X_i)$, $p_{\cdot j} = P(Y = Y_j)$, $i, 1, 2, \dots, r; j = 1, 2, \dots, s$, 则 X 与 Y 的独立性检验就等价于下述检验:

$$H_0: p_{ij} = p_{i\cdot}p_{\cdot j}, \forall 1 \leq i \leq r, 1 \leq j \leq s \longleftrightarrow H_1: \exists (i, j), p_{ij} \neq p_{i\cdot}p_{\cdot j}$$

注: 若 X 与 Y 为连续型随机变量, 这时将它们的取值范围分成 r 个及 s 个互不相交的小区间, 用 n_{ij} 表示 n 个随机试验的结果中 “ X 属于第 i 个小区间, Y 属于第 k 个小区间” 的频数 ($i = 1, 2, \dots, r; k = 1, 2, \dots, s$). 这时可将 X 与 Y 的独立性转化为列联表的独立性检验问题.

由于 $p_{i\cdot}$ 的极大似然估计为 $\hat{p}_{i\cdot} = n_{i\cdot}/n$, $p_{\cdot j}$ 的极大似然估计为 $\hat{p}_{\cdot j} = n_{\cdot j}/n$, 因此若 H_0 成立, 则 p_{ij} 的极大似然估计为 $\hat{p}_{ij} = n_{i\cdot}n_{\cdot j}/n^2$. 从而 X 取 X_i , Y 取 Y_j (试验数据落入第 (i, j) 个类) 的理论频数为 $n \times n_{i\cdot}n_{\cdot j}/n^2 = n_{i\cdot}n_{\cdot j}/n$. 由此构造检验统计量

$$\chi^2 = \sum_{i=1}^r \sum_{k=1}^s \left[n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n} \right]^2 / \frac{n_{i\cdot}n_{\cdot j}}{n} \quad (7-3.1)$$

可以证明在原假设成立时, χ^2 近似服从 $\chi^2((r-1)(s-1))$.

R语言中函数 `chisq.test()` 可完成独立性检验, `chisq.test()` 的调用格式见 §7.2.

例 7.3.1 表 7.4 是对 63 个肺癌患者和由 43 人组成的对照组的调查结果. 问总体中患肺癌是否与吸烟有关系? ($\alpha = 0.05$)

表 7.4 吸烟与肺癌关系的调查数据

	吸烟	不吸烟
肺癌患者	60	3
对照组	32	11

解 R程序如下:

```
> compare<-matrix(c(60,32,3,11), nr = 2,
                    dimnames = list(c("cancer", "normal"),
                                     c("smoke", "Not smoke"))))
> chisq.test(compare, correct=TRUE)
```

运行结果为:

```
Pearson's Chi-squared test with Yates' continuity correction
data:  compare
X-squared = 7.93, df = 1, p-value = 0.004855
```

结论: 因为 p 值 $=0.004855 < \alpha = 0.05$, 故拒绝原假设, 即认为患肺癌与吸烟有关系. ■

7.3.2 Fisher精确检验

上述近似 χ^2 检验要求2维列联表中只允许20%以下的格子的期望频数小于5, 小于R会给出警告, 这时应该使用Fisher精确检验. 下面仅以 2×2 列联表(见表7.5)加以叙述.

在 X 和 Y 独立的原假设下, 在给定边际频率时, 这个具体的列联表的条件概率只依赖于四个值中的任意一个, 其条件概率为:

$$P\{n_{ij}\} = \frac{n_{1.}!n_{2.}!n_{.1}!n_{.2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}, i = 1, 2, j = 1, 2, \quad (7-3.2)$$

即 n_{ij} 服从超几何分布.

表 7.5 2×2 列联表

	B_1	B_2	总和
A_1	n_{11}	n_{12}	$n_{1.}$
A_2	n_{21}	n_{22}	$n_{2.}$
总和	$n_{.1}$	$n_{.2}$	n

在给定 $n_{11} + n_{21} = n_{.1}$ 后, 我们在 n_{11} 比较大时拒绝 H_0 , 所以给定水平 α , 它的临界值 C 满足条件:

$$P(n_{11} \geq C) = \sum_{i \geq C} \frac{n_{1.}! n_{2.}! n_{.1}! n_{.2}!}{n! i! (n_{.1} - i)! (n_{1.} - i)! (n - n_{1.} - n_{.1} - i)!} \leq \alpha.$$

R语言中的`fisher.test()`函数可完成原假设的检验. `fisher.test()`的调用格式如下:

fisher.test() 的调用格式

```
fisher.test(x, y=NULL, workspace=200000, hybrid=FALSE, control=list(),
            or = 1, alternative = "two.sided", conf.int = TRUE,
            conf.level = 0.95, simulate.p.value = FALSE, B = 2000)
```

说明: 参数`workspace`的值为整数, 指定工作空间的数量; 参数`hybrid`的值为逻辑型, 指定是否计算精确的概率, 这两个参数只在维数高于 2×2 的列联表中使用; 参数`or`指定假设的概率比率, 只在 2×2 列联表中使用.

例 7.3.2 数据同例7.3.1, 问总体中肺癌患者吸烟的比例是否比对照组中吸烟的比例要大? ($\alpha = 0.05$)

解 R程序如下:

```
> compare<-matrix(c(60,32,3,11),nr = 2,
                  dimnames = list(c("cancer", "normal"),
                                   c("smoke", "Not smoke"))))
> fisher.test(compare, alternative = "greater")
```

运行结果为:

```

Fisher's Exact Test for Count Data
data:  compare
p-value = 0.002467
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 1.95  Inf
sample estimates:
odds ratio
 6.74691

```

结论: 因为 p 值 $=0.002467 < \alpha = 0.05$, 故拒绝原假设, 认为总体中肺癌患者吸烟的比例是要比对照组中吸烟的比例大. ■

7.3.3 Wilcoxon秩和检验法和Mann-Whitney U检验

Wilcoxon秩和检验法

在正态总体的假定下, 两样本的均值检验通常用 t 检验. 检验统计量

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})S^2}}$$

在零假设成立时服从自由度为 $n_1 + n_2 - 2$ 的 t 分布. 和单样本情况一样, t 检验并不稳健, 在不知总体分布时, 使用 t 检验可能有风险. 这时考虑非参数方法: Wilcoxon秩和检验法.

此检验法是用来检验两个样本的位置参数关系. 与单样本的Wilcoxon符号检验一样, 它也充分利用了样本中秩的信息. 此检验需要的假设:

设 X_1, X_2, \dots, X_m 为来自连续型总体 X 的容量为 m 的样本, Y_1, Y_2, \dots, Y_n 分别来自连续型总体 Y 的容量为 n 的样本, 且两样本相互独立. 记 M_X 为总体 X 的中位数, M_Y 为总体 Y 的中位数.

考虑假设检验问题:

- 1) $H_0 : M_X = M_Y \longleftrightarrow H_1 : M_X > M_Y$ (单边假设检验)
- 2) $H_0 : M_X = M_Y \longleftrightarrow H_1 : M_X < M_Y$ (单边假设检验)
- 3) $H_0 : M_X = M_Y \longleftrightarrow H_1 : M_X \neq M_Y$ (双边假设检验)

构造检验统计量的基本思想是: 把样本 X_1, X_2, \dots, X_m 和 Y_1, Y_2, \dots, Y_n 混合起来, 并把这 $N = (m + n)$ 个观测值从小到大排列起来, 这样每一个 Y 的观察值在混合排列中都有自己的秩. 令 R_i 为 Y_i 在这 N 个数中的秩, 则这些秩的和为 $W_Y = \sum_{i=1}^n R_i$. 同样地由 X 的样本也可得到 W_X , 称 W_X 或 W_Y 为Wilcoxon秩和统计量, 它们的分布由下面的定理给出.

定理 7.3.1 在原假设 H_0 为真时, W_Y 的概率分布和累积概率分别为:

$$\begin{aligned} P(W_Y = d) &= P\left(\sum_{i=1}^n R_i = d\right) = \frac{t_{m,n}(d)}{\binom{N}{n}} \\ P(W_Y \leq d) &= P\left(\sum_{i=1}^n R_i \leq d\right) = \frac{\sum_{i \leq d} t_{m,n}(i)}{\binom{N}{n}} \end{aligned} \quad (7-3.3)$$

其中 $d = \frac{n(n+1)}{2}, \dots, \frac{n(n+1)}{2} + mn$; $t_{m,n}(d)$ 表示 $1, 2, \dots, N = (m + n)$ 这 N 个数中任取 n 个数, 其和恰为 d 的取法数.

由定理??可以给出以上三个假设检验问题的拒绝域及 p 值(略).

另外, 在样本量比较大时, 精确算法的计算量很大. 可以考虑用大样本近似来简化计算和检验. 可以证明在原假设 H_0 为真时,

$$E(W_Y) = \frac{n(N+1)}{2}, \quad \text{Var}(W_Y) = \frac{mn(N+1)}{12},$$

故当 m, n 比较大时,

$$Z = \frac{W_Y - \frac{n(N+1)}{2}}{\sqrt{\frac{mn(N+1)}{12}}} \sim N(0, 1). \quad (7-3.4)$$

Mann-Whitney U 检验

与Wilcoxon秩和统计量等价的有Mann-Whitney U 统计量. 令 W_{XY} 为把所有的 X 的观察值和 Y 的观察值做比较之后, Y 的观察值大于 X 的观察值的个数, 则称 W_{XY} 为Mann-Whitney U 统计量. 它与Wilcoxon秩和统计量的关系如下:

$$W_Y = W_{XY} + \frac{n(n+1)}{2}, \quad W_X = W_{YX} + \frac{m(m+1)}{2}.$$

故可以根据定理7.3.1给出 W_{XY} 的概率分布和累积概率, 从而可以对假设检验问题给出拒绝域和 p 值.

R语言中函数`wilcoxon.test()`可完成原假设的检验, 其调用格式见7.1.2.

例 7.3.3 有糖尿病的和正常的老鼠重量为(单位: 克)

糖尿病鼠: 42, 44, 38, 52, 48, 46, 34, 44, 38;

正常老鼠: 34, 43, 35, 33, 34, 26, 30, 31, 31, 27, 28, 27, 30, 37, 32.

检验这两组的体重是否有显著不同? ($\alpha = 0.05$)

解 R程序如下:

```
> diabetes<-c(42,44,38,52,48,46,34,44,38)
> normal<-c(34,43,35,33,34,26,30,31,31,27,28,27,30,37,32)
> wilcox.test(diabetes,normal,exact = FALSE, correct=FALSE)
```

运行结果为:

```
Wilcoxon rank sum test
data: diabetes and normal
W = 128, p-value = 0.0003008
alternative hypothesis: true location shift is not equal to 0
```

结论: 因为 p 值=0.0003008 < $\alpha = 0.05$, 故拒绝原假设, 认为这两组的体重显著不同. ■

7.3.4 Mood检验

位置参数描述了总体的位置, 而描述总体概率分布离散程度的参数是尺度参数. 假定两独立样本 X_1, X_2, \dots, X_m 和 Y_1, Y_2, \dots, Y_n 分别来自 $N(\mu_1, \sigma_1^2)$ 和 $N(\mu_2, \sigma_2^2)$, 则检验 $H_0: \sigma_1^2 = \sigma_2^2$ 最常用的传统的统计方法是 F 检验, 检验统计量为两独立样本的方差之比 $F = S_X^2/S_Y^2$. 在零假设成立时, 它服从自由度为 $(m-1, n-1)$ 的 F 分布. 但是在总体不是正态或有严重污染时, 上述的 F 检验就不一定合适了. 本小节介绍的Mood检验是用来检验两样本尺度参数之间关系的一种非参数方法.

设两连续总体 X 与 Y 独立, 样本 $X_1, X_2, \dots, X_m \sim F(\frac{x-\theta_1}{\sigma_1})$, $Y_1, Y_2, \dots, Y_n \sim F(\frac{y-\theta_2}{\sigma_2})$, 而且 $F(0) = \frac{1}{2}$, $\theta_1 = \theta_2$. (若不相等, 可以通过平移来使它们相等)

考虑假设检验问题:

$$1) H_0 : \sigma_1 = \sigma_2 \longleftrightarrow H_1 : \sigma_1 > \sigma_2 (\text{单边假设检验})$$

$$2) H_0 : \sigma_1 = \sigma_2 \longleftrightarrow H_1 : \sigma_1 < \sigma_2 (\text{单边假设检验})$$

$$3) H_0 : \sigma_1 = \sigma_2 \longleftrightarrow H_1 : \sigma_1 \neq \sigma_2 (\text{双边假设检验})$$

构造检验统计量的基本思想为: 把样本 X_1, X_2, \dots, X_m 和 Y_1, Y_2, \dots, Y_n 混合起来, 记 $R_{11}, R_{12}, \dots, R_{1m}$ 为 X 的观察值在混合样本中的秩, 而 $R_{21}, R_{22}, \dots, R_{2n}$ 为 Y 的观察值在混合样本中的秩, $N = m + n$. 对样本 X 来说, 考虑秩统计量

$$M = \sum_{j=1}^m \left(R_{1j} - \frac{N+1}{2} \right)^2. \quad (7-3.5)$$

则以上三个假设检验问题的拒绝域分别为:

$$C_1 = \{M \geq c\}, \text{ 其中 } c \text{ 满足: } c = \inf\{c^* : P(M \geq c^*) \leq \alpha\};$$

$$C_2 = \{M \leq d\}, \text{ 其中 } d \text{ 满足: } d = \sup\{d^* : P(M \leq d^*) \leq \alpha\};$$

$$C_3 = \{M \geq c \text{ 或 } M \leq d\}, \text{ 其中 } c, d \text{ 满足:}$$

$$c = \inf\{c^* : P(M \geq c^*) \leq \frac{\alpha}{2}\}, \quad d = \sup\{d^* : P(M \leq d^*) \leq \frac{\alpha}{2}\}. \quad (7-3.6)$$

当原假设 H_0 成立时, 可以证明:

$$\begin{aligned} E(M) &= \frac{m(N^2 - 1)}{12}, \\ \text{Var}(M) &= \frac{mn}{N(N-1)} \sum_{i=1}^N \left[\left(i - \frac{N+1}{2} \right)^2 - \frac{N^2 - 1}{12} \right]^2. \end{aligned} \quad (7-3.7)$$

故

$$Z = \frac{M - E(M)}{\sqrt{\text{Var}(M)}} \sim N(0, 1). \quad (7-3.8)$$

R语言中函数`mood.test()`可完成原假设的检验, 其调用格式如下

mood.test()的调用格式

```
mood.test(x, y, alternative =  
          c("two.sided", "less", "greater"),...)
```

例 7.3.4 两个村农民的月收入分别为(单位: 元)

A村: 321, 266, 256, 388, 330, 329, 303, 334, 299, 221, 365, 250, 258, 342, 343, 298, 238, 317, 354;

B村: 488, 598, 507, 428, 807, 342, 512, 350, 672, 589, 665, 549, 451, 481, 514, 391, 366, 468.

问两个村农民的月收入的内部差异是否相同? ($\alpha = 0.05$)

解 R程序如下:

```
> A<-c(321, 266, 256, 388, 330, 329, 303, 334, 299,  
       221, 365, 250, 258, 342, 343, 298, 238, 317, 354)  
> B<-c(488, 598, 507, 428, 807, 342, 512, 350, 672,  
       589, 665, 549, 451, 481, 514, 391, 366, 468)  
> diff<-median(B)-median(A)  
> A<-A+diff  
> mood.test(A,B)
```

运行结果为:

```
Mood two-sample test of scale  
data:  A and B  
Z = -2.4846, p-value = 0.01297  
alternative hypothesis: two.sided
```

结论: 因为 p 值 $=0.01297 < \alpha = 0.05$, 故拒绝原假设, 认为这两个村的内部差异是不同的. ■

注意: 因为mood检验需要的假定之一是要两样本的中位数相同, 故在做检验时, 要先消除两样本之间中位数的差异, 接着才可以做mood检验.

§7.4 多总体的比较与检验

多样本问题是统计中最常见的一类问题. 例如多种投资方案在试行后效果的比较、不同机器在同一条件下的稳定性是否相同等等. 本节就多样本模型讨论位置参数与尺度参数的检验问题.

设 k 个连续型随机变量(总体) X_1, X_2, \dots, X_k 相互独立, $X_i \sim F\left(\frac{x-\theta_i}{\sigma_i}\right), \sigma_i > 0$, $X_{i1}, X_{i2}, \dots, X_{in_i}$ 是来自第 i 个总体 X_i 的容量为 n_i 的样本, $N = \sum_{i=1}^k n_i$.

7.4.1 位置参数的Kruskal-Wallis秩和检验

设 $\sigma_1 = \sigma_2 = \dots = \sigma_k$, 不妨设为1 (其检验见下面二小节). 考虑假设检验问题:

$$H_0: \theta_1 = \theta_2 = \dots = \theta_k \longleftrightarrow H_1: \theta_1, \theta_2, \dots, \theta_k \text{不全相等}$$

构造检验统计量的基本思想为: 把 k 个样本混合起来, 算出所有数据在混合样本中的秩, 记样本 X_{ij} 的秩为 R_{ij} (R_{ij} 的意义同§7.3.4), 对每一个样本的观察值的秩求和得到 $R_i = \sum_{j=1}^{n_i} R_{ij}$, $i = 1, 2, \dots, k$, 由此找到它们在每组中的平均值 $\bar{R}_i = R_i/n_i$. 如果这些 \bar{R}_i 很不一样, 就可以怀疑原假设.

构造检验统计量:

$$\begin{aligned} H &= \frac{12}{N(N+1)} \sum_{i=1}^k n_i (\bar{R}_i - \bar{R})^2 \\ &= \frac{12}{N(N+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(N+1) \end{aligned} \quad (7-4.1)$$

其中 $\bar{R} = \sum_{i=1}^k n_i \bar{R}_i / N = \frac{N+1}{2}$.

可以证明:

$$E(R_i) = n_i(N+1), \text{Var}(R_i) = \frac{n_i(N-n_i)(N+1)}{12}.$$

从而

$$\begin{aligned}
 E(\bar{R}_i) &= N + 1, \\
 \text{Var}(\bar{R}_i) &= \frac{(N - n_i)(N + 1)}{12n_i}, \\
 E(H) &= \frac{12}{N(N + 1)} E\left(\sum_{i=1}^k n_i \left(\bar{R}_i - \frac{N + 1}{2}\right)^2\right) \\
 &= \frac{12}{N(N + 1)} \sum_{i=1}^k n_i \text{Var}(\bar{R}_i) = k - 1. \quad (7-4.2)
 \end{aligned}$$

当原假设 H_0 成立时, 若 $\min\{n_1, n_2, \dots, n_k\} \rightarrow +\infty$, 且 $\frac{n_i}{N} \rightarrow \lambda_i$, $i = 1, 2, \dots, k$, $\lambda_i \in (0, 1)$, 则 $H \sim \chi^2(k - 1)$.

故上述检验问题的拒绝域 $C = \{H \geq \chi_{1-\alpha}^2(k - 1)\}$.

R中函数`kruskal.test()`可完成原假设的检验其调用格式如下:

————— `kruskal.test()` 的调用格式 —————

`kruskal.test(x, g, ...)`

说明: x 为一向量或列表, g 为对 x 分类的因子, 当 x 为列表时 g 可以省略.

例 7.4.1 下面的数据是游泳、打篮球、骑自行车等三种不同的运动在30分钟内消耗的热量(单位:卡路里). 这些数据是否说明这三种运动消耗的热量全相等? ($\alpha = 0.05$)

游泳: 306, 385, 300, 319, 320;

打篮球: 311, 364, 315, 338, 398;

骑自行车: 289, 198, 201, 302, 289.

解 **R**程序如下:

```
> x<-list(swim=c(306, 385, 300, 319, 320),
          basketball=c(311, 364, 315, 338, 398),
          bicycle=c(289, 198, 201, 302, 289))
> kruskal.test(x)
```

运行结果为:

```
Kruskal-Wallis rank sum test
data: x
Kruskal-Wallis chi-squared = 9.1564, df = 2, p-value = 0.01027
```

结论: 因为 p 值 $=0.01027 < \alpha = 0.05$, 故拒绝原假设, 认为这三种运动消耗的热量不全相等. ■

7.4.2 尺度参数的Ansari-Bradley检验

设 $\theta_1 = \theta_2 = \cdots = \theta_k$. 考虑假设检验问题:

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 \longleftrightarrow H_1: \sigma_1^2, \sigma_2^2, \cdots, \sigma_k^2 \text{不全相等}$$

记

$$\bar{A}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \left[\frac{N+1}{2} - |R_{ij} - \frac{N+1}{2}| \right]^2, i = 1, 2, \cdots, k$$

构造检验统计量:

$$B = \frac{N^3 - 4N}{48(N+1)} \sum_{i=1}^k n_i \left[\bar{A}_i - \frac{N+2}{4} \right]^2. \quad (7-4.3)$$

可以证明在原假设 H_0 成立时, $B \sim \chi^2(k-1)$. 从而上述检验问题的拒绝域为 $C = \{B \geq \chi_{1-\alpha}^2(k-1)\}$.

R语言中函数`ansari.test()`可完成原假设的检验, 其调用格式如下:

—— `ansari.test()` 的调用格式 ——

```
ansari.test(x, y, alternative = c("two.sided", "less", "greater"),
            exact = NULL, conf.int = FALSE, conf.level = 0.95, ...)
```

说明: x 为一向量或列表, g 为对 x 分类的因子, 当 x 为列表时 g 可以省略.

例 7.4.2 两个工人加工的零件尺寸(各10个)为(单位:mm):

工人A: 18.0, 17.1, 16.4, 16.9, 16.9, 16.7, 16.7, 17.2, 17.5, 16.9;

工人B: 17.0, 16.9, 17.0, 16.9, 17.2, 17.1, 16.8, 17.1, 17.1, 17.2.

这个结果能否说明两个工人的水平(加工精度)一致? $(\alpha = 0.05)$

解 R程序如下:

```
> worker.a<-c(18.0,17.1,16.4,16.9,16.9,16.7,16.7,17.2,17.5,16.9)
> worker.b<-c(17.0,16.9,17.0,16.9,17.2,17.1,16.8,17.1,17.1,17.2)
> ansari.test(worker.a,worker.b)
```

运行结果为:

```
Ansari-Bradley test
data: worker.a and worker.b
AB = 41.5, p-value = 0.04232
alternative hypothesis: true ratio of scales is not equal to 1
Warning message:
In ansari.test.default(worker.a, worker.b) :
cannot compute exact p-value with ties
```

结论: 因为 p 值 $=0.04232 < \alpha = 0.05$, 故拒绝原假设, 认为这两个工人的水平不一样. 最后一句的警告信息是因为原数据中有结. 其定义及处理方法见附录A.

■

7.4.3 尺度参数的Fligner-Killeen检验

该检验需要的假设与同Ansari-Bradley检验.

记 $\theta_1 = \theta_2 = \cdots = \theta_k = \theta$, $V_{ij} = |X_{ij} - \theta|$, $i = 1, 2, \cdots, k$; $j = 1, 2, \cdots, n_i$. 当 θ 未知时, 用样本中位数 M 代替 θ , 即 $V_{ij} = |X_{ij} - M|$, 再用 R_{ij} 表示在混合样本中 V_{ij} 的秩.

$k=2$ 时, 采用检验统计量

$$W = \sum_{i=1}^{n_1} R_{ij}.$$

可以证明在原假设 H_0 成立时, 统计量 W 有Wilcoxon分布;

$k > 2$ 时, 采用检验统计量

$$K = \frac{12}{N(N+1)} \sum_{i=1}^k n_i \left(\bar{R}_i - \frac{N+1}{2} \right)^2,$$

其中 $\bar{R}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} R_{ij}$. 可以证明在 H_0 成立时, 统计量 K 有Kruskal-Wallis零分

布.

R语言中函数`fligner.test()`可完成原假设的检验, 其调用格式如下:

fligner.test() 的调用格式

```
fligner.test(x, g, ...)
```

说明: x 为一向量或列表, g 为对 x 分类的因子, 当 x 为列表时 g 可以省略.

例 7.4.3 三名不同的运动员A、B、C同时在同一条件下进行打靶比赛, 各打10发子弹, 他们打中的环数如下:

A: 8, 7, 9, 10, 9, 6, 5, 8, 10, 5;

B: 8, 7, 9, 6, 8, 9, 10, 7, 8, 9;

C: 10, 10, 9, 6, 8, 3, 5, 6, 7, 4.

问这三名运动员的稳定性是否一样? ($\alpha = 0.05$)

解 R程序如下:

```
> x<-list(A=c(8,7,9,10,9,6,5,8,10,5),
          B=c(8,7,9,6,8,9,10,7,8,9),
          C=c(10,10,9,6,8,3,5,6,7,4))
> fligner.test(x)
```

运行结果为:

```
Fligner-Killeen test of homogeneity of variances
data:  x
Fligner-Killeen:med chi-squared = 5.1905, df = 2, p-value =0.07463
```

结论: 因为 p 值 $=0.07463 > \alpha = 0.05$, 故接受原假设, 认为这三名运动员的稳定性相同. ■

第七章习题

7.1 某地区从事管理工作的职员月收入的中位数是6500元. 现有一个该地区从事管理工作的20个妇女组成的样本, 她们的月收入如下:

6100, 5300, 4900, 7100, 6400, 5700, 5200, 5100, 6800, 6200, 7000, 3900, 5300, 6200, 6500, 6300, 6200, 5300, 5800, 6700.

问该地区从事管理工作的妇女月收入的中位数是否小于6500? ($\alpha = 0.05$)

7.2 调查某美发店上半年各月顾客数量, 如表7.6所示. 问该店每月的顾客数量是否服从均匀分布?

表 7.6 美发店1—6月份顾客数量

月份	1	2	3	4	5	6	合计
顾客人数(百人)	27	18	15	24	36	30	150

7.3 从某地区高中二年级学生中随机抽取45位学生测得他们的体重如表7.7所示, 问该地区学生的体重是否服从正态分布?

表 7.7 高二年级学生体重(单位: 公斤)

36	36	37	38	40	42	43	43	44	45	48	48	50	50	51
52	53	54	54	56	57	57	57	58	58	58	58	58	59	60
61	61	61	62	62	63	63	65	66	68	68	70	73	73	75

7.4 美国某年总统选举前, 由社会调查总部抽查黑白种族与支持不同政党是否有关, 得到数据如表7.8所示, 问不同种族与支持持政党之间是否存在独立性? ($\alpha = 0.05$)

7.5 为了解两种药物对治疗某种疾病的效果, 抽取42名患者分别服用药物A和B, 数据如表所示, 问药物的疗效与服用的药物是否相关? ($\alpha = 0.05$)

7.6 在一次社会调查中, 以问卷的方式调查了总共901人的年收入及对工作的满意程度, 其中年收入(A)分为小于6000元、6000 ~ 15000元、15000 ~

表 7.8 种族与政党的关系数据

种族	民主党	共和党	无党
白人	341	405	105
黑人	103	11	15

表 7.9 某疾病两种药物的治疗效果

药物	疗效		
	有效	无效	合计
A	8	2	10
B	34	18	32
合计	22	20	42

25000元及超过25000元4档. 对工作的满意程度(B)分为很不满意、较不满意、基本满意和很满意4档. 调查结果如表7.10所示. 问工作的满意程度与年收入高低是否无关? ($\alpha = 0.05$)

表 7.10 工作满意程度与年收入列联表

	很不满意	较不满意	基本满意	很满意	合计
< 6000	20	24	80	82	206
6000 ~ 15000	22	38	104	125	289
15000 ~ 25000	13	28	81	113	235
> 25000	7	18	54	92	171
合计	62	108	319	412	901

7.7 股票的波动程度可以用来衡量投资的风险. 取自同一年11月和12月的前10个交易日的股票指数样本数据, 如下:

11月: 1149, 1169, 1152, 1183, 1173, 1169, 1130, 1152, 1120, 1171;

12月: 1116, 1147, 1135, 1125, 1184, 1125, 1192, 1174, 1164, 1180.

问:

- 1) 这两段时间的股票指数的中位数是否相同? $(\alpha = 0.05)$
- 2) 这两段时间的股票指数的波动程度是否一样? $(\alpha = 0.05)$

7.8 对5位健康成年人的血液测量其中的尿酸浓度, 分别用手工(X)和仪器(Y)两种方法测量, 结果如表7.11所示, 问两种测量方法的精度是否存在差异? $(\alpha = 0.05)$

表 7.11 尿酸浓度的两种测量值

手工(X)	4.5	6.5	7	10	12
仪器(Y)	6	7.2	8	9	9.8

7.9 茶是世界上最为广泛的一种饮料, 但是很少人知其营养价值. 任一种茶叶都含有叶酸, 它是一种维他命B. 如今已有测定茶叶中叶酸含量的方法. 为研究各产地的绿茶的叶酸含量是否有显著差异, 特选四个产地绿茶, 其中A制作了7个样品, B制作了5个样品, C和D各制作了6个样品, 共有24个样品, 按随机次序测试其叶酸含量(单位:mg), 测试结果如表7.12所示.

表 7.12 四个产地茶叶的叶酸含量

产地	叶酸含量(单位:mg)
A	7.9, 6.2, 6.6, 8.6, 10.1, 9.6, 8.9
B	5.7, 7.5, 9.8, 6.1, 8.4
C	6.4, 7.1, 7.9, 4.5, 5.0, 4.0
D	6.8, 7.5, 5.0, 5.3, 6.1, 7.4

问:

- 1) 四个产地绿茶的叶酸含量的均值是否有显著差异? $(\alpha = 0.05)$
- 2) 四个产地绿茶的叶酸含量的方差是否有显著差异? $(\alpha = 0.05)$

第八章 方差分析

本章概要

- ◇ 单因子方差分析
- ◇ 双因子方差分析
- ◇ 协方差分析

方差分析(analysis of variance, 简称为ANOVA)是工农业生产和科学研究中分析试验数据的一种有效的统计方法. 引起观测值不同(波动)的原因主要有两类: 一类是试验过程中随机因素的干扰或观测误差所引起不可控制的波动, 另一类则是由于试验中处理方式不同或试验条件不同引起的可以控制的波动. 方差分析的主要工作就是将观测数据的总变异(波动)按照变异的原因的不同分解为因子效应与试验误差, 并对其作出数量分析, 比较各种原因在总变异中所占的重要程度, 以此作为进一步统计推断的依据.

§8.1 单因子方差分析

8.1.1 数学模型

设试验只有一个因子(又称为因素) A 有 r 个水平 A_1, A_2, \dots, A_r . 现在水平 A_i 下进行 n_i 次独立观测, 得到观测数据为 $X_{ij}, j = 1, 2, \dots, n_i, i = 1, 2, \dots, r$, 则单因素方差模型可表示为

$$\begin{cases} X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, i = 1, 2, \dots, r, j = 1, 2, \dots, n_i, \\ \varepsilon_{ij} \sim N(0, \sigma^2), \text{且各}\varepsilon_{ij}\text{相互独立}, \\ \sum_{i=1}^r n_i \alpha_i = 0. \end{cases} \quad (8-1.1)$$

其中 μ 为总平均, α_i 是第 i 个水平的效应, ε_{ij} 是随机误差. 若 $n_1 = n_2 = \cdots = n_r$, 称模型是平衡的, 否则称为非平衡的.

我们的目的是要比较因素 A 的 r 个水平的京郊是否有显著差异, 这可归结为检验假设

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_r \longleftrightarrow H_1: \alpha_1, \alpha_2, \dots, \alpha_r \text{ 不全相等}$$

如果 H_0 被拒绝, 则说明因素 A 的各水平的效应之间有显著的差异, 否则, 差异不明显.

按照方差分析的思想, 将总离差平方和分解为二部分, 即

$$SS_T = SS_E + SS_A$$

其中

$$\begin{aligned} SS_T &= \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2, & \bar{X} &= \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} X_{ij} \\ SS_E &= \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2, & \bar{X}_{i.} &= \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \\ SS_A &= \sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{X}_{i.} - \bar{X})^2 \end{aligned}$$

这里称 SS_T 为总离差平方和(或称总变差), 它是所有数据 X_{ij} 与总平均值 \bar{X} 之差的平方和, 描绘所有观察数据的离散程度; SS_E 为误差平方和(或组内平方和), 是对固定的 i , 观测值 $X_{i1}, X_{i2}, \dots, X_{in_i}$ 之间的差异大小的度量. SS_A 为因素 A 的效应平方(和或组间平方和), 表示因子 A 各水平下的样本均值和总平均值之差的平方和.

可以证明, 当 H_0 成立时

$$\frac{SS_E}{\sigma^2} \sim \chi^2(n-r), \quad \frac{SS_A}{\sigma^2} \sim \chi^2(r-1),$$

且 SS_A 与 SS_E 独立. 于是

$$F = \frac{SS_A/(r-1)}{SS_E/(n-r)} \sim F(r-1, n-r) \quad (8-1.2)$$

若 $F > F_{\alpha}(r-1, n-r)$, 则拒绝原假设, 认为因素A的 r 个水平有显著差异, 反之“接受”原假设. 这也可以通过检验的 p 值来决定是接受还是拒绝原假设 H_0 .

R中函数aov()提供了方差分析的计算与检验, 其调用格式为

aov()的调用格式

```
aov(formula, data=NULL, projections=FALSE,
     qr=TRUE, contrasts=NULL, ...)
```

说明: formula是方差分析的公式, 在单因素方差分析中它表示为 $x \sim A$, data是数据框, 其它参见在线帮助.

例 8.1.1 以淀粉为原料生产葡萄糖的过程中, 残留许多糖蜜, 可作为生产酱色的原料. 在生产酱色的过程之前应尽可能彻底除杂, 以保证酱色质量. 为此对除杂方法进行选择. 在实验中选用5种不同的除杂方法, 每种方法做4次试验, 即重复4次, 结果见表8.1.

表 8.1 不同除杂方法的除杂量

除杂方法 A_i	除杂量 X_{ij}				均量 \bar{X}_i
A_1	25.6	22.2	28.0	29.8	26.4
A_2	24.4	30.0	29.0	27.5	27.7
A_3	25.0	27.7	23.0	32.2	27.0
A_4	28.8	28.0	31.5	25.9	28.6
A_5	20.6	21.2	22.0	21.2	21.3

解 **R**程序为:

```
> X<-c(25.6, 22.2, 28.0, 29.8, 24.4, 30.0, 29.0, 27.5, 25.0, 27.7,
       23.0, 32.2, 28.8, 28.0, 31.5, 25.9, 20.6, 21.2, 22.0, 21.2)
> A<-factor(rep(1:5, each=4))
> miscellany<-data.frame(X, A)
> aov.mis<-aov(X~A, data=miscellany)
> summary(aov.mis)
```

输出结果为

```

      Df  Sum Sq Mean Sq F value  Pr(>F)
A         4 131.957   32.989   4.3061 0.01618 *
Residuals 15 114.915    7.661
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

■

说明: 上述结果中, *Df*表示自由度; *sum Sq*表示平方和; *Mean Sq*表示均方和; *F value*表示*F*检验统计量的值, 即*F*比; *Pr(>F)*表示检验的*p*值; *A*就是因素*A*; *Residuals*为残差.

可以看出, $F = 4.3061 > F_{0.05}(5-1, 20-5) = 3.06$, 或者 $p=0.01618<0.05$, 说明有理由拒绝原假设, 即认为五种除杂方法有显著差异. 据上述结果可以填写下面的方差分析表: 再通过函数*plot()*绘图可直观描述5种不同除杂方法之

表 8.2 除杂方法试验的方差分析表

方差来源	自由度	平方和	均方和	<i>F</i> 比	<i>p</i> 值
因素A	4	131.957	32.989	4.3061	0.01618
误差	15	114.915	7.661		
总和	19	246.872			

间的差异, **R**中运行命令

```
> plot(miscellany$X~miscellany$A)
```

得到图8.1. 从图形上也可以看出, 5种除杂方法产生的除杂量有显著差异, 特别第5种与前面的4种, 而方法1与3, 方法2与4的差异不明显.

8.1.2 均值的多重比较

进行方差分析后发现各效应的均值之间有显著差异, 此时只能知道有某些均值彼此不同, 但无法知道哪些均值不同, 下面的方法帮助我们找出在进行方差分析时哪些均值是不同的.

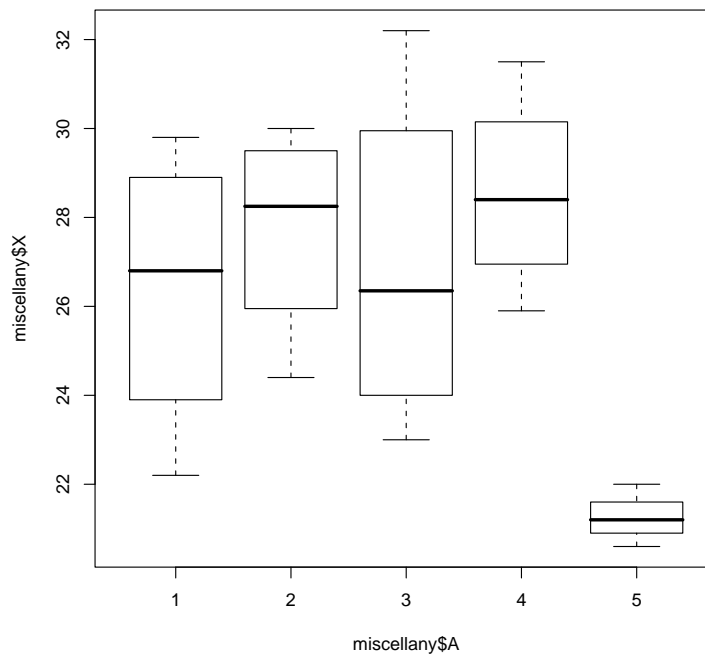


图 8.1 不同除杂方法的差异

多重t检验方法

这种方法就是针对因子A的两个效应进行比较, 假设检验为

$$H_0: \alpha_i = \alpha_j, i \neq j, (i, j = 1, 2, \dots, r)$$

检验统计量为

$$T_{ij} = \frac{\bar{X}_{i\cdot} - \bar{X}_{j\cdot}}{\sqrt{MS_E(\frac{1}{n_i} + \frac{1}{n_j})}}, \quad i \neq j, i(j = 1, 2, \dots, r)$$

其中 $MS_E = SS_E/(n - r)$ 为误差的均方和, 也是 σ^2 的估计. 当 H_0 成立时, $T_{ij} \sim t(n - r)$. 所以检验的拒绝域为

$$C = \{|T_{ij}| > t_{1-\frac{\alpha}{2}}(n - r)\}. \quad (8-1.3)$$

说明: 多重 t 检验方法使用方便, 但当多次重复使用 t 检验时会增大犯第一类错误的概率, 从而使得“有显著差异”的结论不一定可靠, 所以在进行较多次重复比较时, 我们要对 p 值进行调整.

R软件中 p 值调整使用函数`p.adjust()`, 其调用格式为

`p.adjust()`的调用格式

```
p.adjust(p, method=p.adjust.methods, n=length(p))
```

说明: p 是 p 值构成的向量, `method`是修正方法, 包括

- Holm(1979)方法
- Hochberg(1988)方法
- Hommel(1988)方法
- Bonferroni方法
- Benjamini & Hochberg, BH(1995)方法
- Benjamini & Yekutieli, BY(2001)方法

R中键入命令

```
> p.adjust.methods
```

得到调整方法的列表:

```
[1] "holm"          "hochberg"      "hommel"        "bonferroni"    "BH"
[6] "BY"            "fdr"           "none"
```

具体意义参见在线帮助.

当比较次数较多时, Bonferroni方法的效果较好, 所以在作多重 t 检验时常采用Bonferroni法对 p 进行调整. 实际上, 它采用 $\alpha = \alpha'/k$ 作为给出“有无显著差异”的检验水平, 其中 k 为两两比较的次数, α' 为累积I类错误的概率.

R软件中函数`pairwise.t.test()`可以得到多重比较的 p 值, 其调用格式为

```
pairwise.t.test(x, g, p.adjust.method=p.adjust.methods,
               pool.sd=TRUE, ...)
```

说明: x 是响应变量构成的向量, g 是分组向量(因子). `p.adjust.method` 是上面提到的调整 p 值的方法, “`p.adjust.method=none`” 表示不作任何调整, 默认值按 Holm 方法调整.

例 8.1.2 对例 8.1.1 作均值的多重比较, 进一步检验

$$H_0: \alpha_i = \alpha_j \quad i, j = 1, 2, 3, 4, 5$$

解 用三种方法进行多重比较:

- 不对 p 作出调整: **R** 程序为

```
> pairwise.t.test(X, A, p.adjust.method="none")
```

检验结果如下:

```
data: X and A
```

```
      1      2      3      4
2 0.5087 -      -      -
3 0.7729 0.7069 -      -
4 0.2893 0.6793 0.4335 -
5 0.0189 0.0048 0.0104 0.0020
```

```
P value adjustment method: none
```

检验的结果与图 8.1 一致, 即 μ_5 与其它 4 个差异明显, 后者差异不明显.

- 按缺省的 “holm” 对 p 值进行调整: **R** 程序为

```
> pairwise.t.test(X, A, p.adjust.method="holm")
```

检验结果如下:

```
Pairwise comparisons using t tests with pooled SD
```

```
data: X and A
```

```

      1      2      3      4
2 1.000 -      -      -
3 1.000 1.000 -      -
4 1.000 1.000 1.000 -
5 0.132 0.043 0.084 0.020
```

```
P value adjustment method: holm
```

- 按缺省的“holm”对 p 值进行调整: **R**程序为

```
> pairwise.t.test(X, A, p.adjust.method="bonferroni")
```

检验结果如下:

```
Pairwise comparisons using t tests with pooled SD
```

```
data: X and A
```

```

      1      2      3      4
2 1.000 -      -      -
3 1.000 1.000 -      -
4 1.000 1.000 1.000 -
5 0.189 0.048 0.104 0.020
```

```
P value adjustment method: bonferroni
```

从输出结果可以看出, 作调整后 p 值增大, 在一定程度上克服了多重 t 检验的缺点.



8.1.3 同时置信区间: Tukey法

若经前面的 F 检验, $H_0: \alpha_1 = \cdots = \alpha_r$ 被拒绝了, 则因子 A 的 r 个水平的效应不全相等, 这时我们希望对效应之差 $\alpha_i - \alpha_j$ ($i \neq j$)作出置信区间, 由此了

解哪一些效应不相等. 这里仅介绍一种基于学生化极差分布的TUKEY方法. 这是J.W.Tukey(1952)提出的一种多重比较方法, 是以试验错误率为标准的, 又称真正显著差(honesty significant difference, HSD)法. 该方法基于下面的定理:

定理 8.1.1 设 X_1, X_2, \dots, X_n 是iid的 $N(\mu, \sigma^2)$, $U = m \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(m)$, 且 $U, X_1, \dots, Y - n$ 相互独立, 则

- 1) $\frac{\max_i X_i - \min_i X_i}{\hat{\sigma}^2 / \sigma^2} \sim q(n, m)$, 其中 $q(n, m)$ 表示参数为 n, m 的学生化极差分布.
- 2) 所有 $\alpha_i - \alpha_j, i \neq j$ 的置信系数为 $1 - \alpha$ 的同时置信区间为

$$X_i - X_j - q_{1-\alpha}(n, m) \hat{\sigma} \leq \alpha_i - \alpha_j \leq X_i - X_j + q_{1-\alpha}(n, m) \hat{\sigma}.$$

对于平衡的方差分析模型, 设 $n_1 = \dots = n_r = n, N = nr$, 由

$$\bar{X} \sim N(\mu + \alpha_i, \sigma^2/n)$$

且 \bar{X}_i 与 \bar{X}_j 独立,

$$(N - r) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(N - r),$$

故由定理知, 对一切 $i \neq j$, $\alpha_i - \alpha_j$ 的置信系数为 $1 - \alpha$ 的同时置信区间(称为Turkey区间)为

$$\bar{X}_i - \bar{X}_j \pm q_{1-\alpha}(r, r(n-1)) \frac{\sigma}{\sqrt{n}}.$$

若 $n_i \neq n_j$, 则 $\alpha_i - \alpha_j$ 的置信系数为 $1 - \alpha$ 的同时置信区间近似为

$$\bar{X}_i - \bar{X}_j \pm q_{1-\alpha}(r, r(n-1)) \frac{\sigma}{\sqrt{2}} \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}.$$

在R软件中, 函数`qtukey()`用于计算 q 分位数, 函数`TukeyHSD()`用于计算同时置信区间, 其调用格式为

TukeyHSD() 的调用格式

`TukeyHSD(x, which, ordered=FALSE, conf.level=0.95...)`

说明: x 为方差分析的对象, `which`是给出需要计算比较区间的因子向量, `ordered`是逻辑值, 如果为"true", 则因子的水平先递增排序, 从而使得因子间差异均以正值出现. `conf.level`是置信水平.

例 8.1.3 某商店以各自的销售方式卖出新型手表, 连续四天手表的销售量如表8.3所示, 试考察销售方式之间是否有显著差异.

表 8.3 销售方式与销售量数据表

销售方式	销售量数据			
A_1	23	19	21	13
A_2	24	25	28	27
A_3	20	18	19	15
A_4	22	25	26	23
A_5	24	23	26	27

解 首先以数据框形式生成数据sales.

```
> sales<-data.frame(
  X=c(23, 19, 21, 13, 24, 25, 28, 27, 20, 18,
      19, 15, 22, 25, 26, 23, 24, 23, 26, 27),
  A=factor(rep(1:5, c(4, 4, 4, 4, 4)))
)
```

其次进行方差分析, 由R命令

```
> summary(aov(X~A, sales))
```

得

```
          Df Sum Sq Mean Sq F value    Pr(>F)
A           4 212.800   53.200     7.98 0.001178 **
Residuals   15 100.000    6.667
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
```

可见不同的销售方式有差异.

最后再求均值之差的同时置信区间. R命令为

```
> TukeyHSD(aov(X~A, sales))
```

运行结果为

```
Tukey multiple comparisons of means
 95% family-wise confidence level

Fit: aov(formula = X ~ A, data = sales)

$A
      diff      lwr      upr    p adj
2-1      7  1.362247 12.637753 0.0120117
3-1     -1 -6.637753  4.637753 0.9805632
4-1      5 -0.637753 10.637753 0.0944731
5-1      6  0.362247 11.637753 0.0344328
3-2     -8 -13.637753 -2.362247 0.0041527
4-2     -2 -7.637753  3.637753 0.8062057
5-2     -1 -6.637753  4.637753 0.9805632
4-3      6  0.362247 11.637753 0.0344328
5-3      7  1.362247 12.637753 0.0120117
5-4      1 -4.637753  6.637753 0.9805632
```

可以看出, 共有10个两两比较的结果, $A_3 - A_1$ 、 $A_4 - A_2$ 、 $A_5 - A_2$ 和 $A_5 - A_4$ 的差异是显著的, 其它两两比较的结果均是不显著的. ■

8.1.4 方差齐性检验

前面已提到要进行方差分析, 应具备以下三个条件: (1)可加性, (2)独立正态性, (3)方差齐性. 方差齐性检验就是检验数据在不同水平下方差是否相同. 最常用的方法就是Bartlett检验和Levene检验.

Bartlett检验

方差齐性检验就是检验数据在不同水平下方差是否相同, 方差齐性检验最常用的方法是Bartlett检验和Levene检验. 检验问题为:

H_0 : 各因子水平下的方差相同 $\longleftrightarrow H_1$: 各因子水平下的方差不齐

当处理组的数据较多时, 令 $N = \sum_{i=1}^r n_i$,

$$\begin{aligned} S_i^2 &= \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 \\ S_c^2 &= \frac{\sum_{i=1}^k [\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)]}{\sum_{i=1}^k (n_i - 1)} \\ &= \frac{1}{N - r} \sum_{i=1}^r (n_i - 1) S_i^2 = MS_E \\ C &= 1 + \frac{1}{3(r-1)} \left[\sum_{i=1}^r \frac{1}{n_i - 1} - \frac{1}{N - r} \right] \end{aligned}$$

则在原假设成立下, 统计量

$$\chi^2 = \frac{2.3026}{C} \left[(N - r) \ln S_c^2 - \sum_{i=1}^r (n_i - 1) \ln S_i^2 \right], \quad \nu = k - 1$$

近似服从自由度为 $(r - 1)$ 的 χ^2 分布. 因此对于给定的显著性水平 α , 若 p 值小于 α , 则拒绝 H_0 , 即认为至少有两个水平下的数据的方差不相等; 否则认为数据满足方差齐性的要求.

R软件中, 函数 `Barlett.test()` 提供 Bartlett 检验, 其调用格式为:

Barlett.test() 调用格式

```
bartlett.test(x, g, ...)
bartlett.test(formula, data, subset, no.action, ...)
```

说明: `x` 是由数据构成的向量或列表; `g` 是由因子构成的向量, 当 `x` 是列表时, 此项无效; `formula` 是方差分析公式, `data` 是数据框, 其余参数见在线帮助.

Levene 检验

将原样本观察值作离均差变换, 或离均差平方变换, 然后进行方差分析, 其检验结果用于判断方差是否齐性.

$$(1) d_{ij} = |X_{ij} - \bar{X}_i|; \quad (2) d_{ij} = |X_{ij} - md_i|; \quad (3) d_{ij} = |X_{ij} - \bar{X}_i|^2$$

其中 md_i 为第 i 水平下数据的样本中位数.

Levene检验对原始数据是否为正态不灵敏, 所以比较稳健, 因此推荐采用LEvene方差齐性检验.

R的程序包car中提供了Levene检验的函数`levene.test()`, 其调用格式为:

`levene.test()` 调用格式

`levene.test(x, group)`

说明: `x`是由数据构成的向量, `g`是由因子构成的向量.

例 8.1.4 对例8.1.3的数据作方差齐性检验. 分别用Bartlett检验和levene检验检验方差的齐性.

解 先用Bartlett检验, 程序

```
> bartlett.test(X~A, data=sales)
```

得检验结果:

```
Bartlett test of homogeneity of variances
```

```
data: X by A
```

```
Bartlett's K-squared = 3.7231, df = 4, p-value = 0.4448
```

即 p 值(0.4448)>0.05, 接受原假设, 认为各处理组的数据是等方差的.

再用levene检验, 程序

```
> library(car)
```

```
> levene.test(sales$X, sales$A)
```

得检验结果:

```
Levene's Test for Homogeneity of Variance
```

```
  Df F value Pr(>F)
```

```
group 4  0.8182 0.5333
```

```
15
```

即 p 值(0.5333)>0.05, 接受原假设, 认为各处理组的数据满足方差齐性的要求. 因此两种检验方法有完全相同的结果. ■

注:

- 1) 方差分析模型可视为一种特殊的线性模型, 因此方差分析还可以使用第九章讲的线性模型函数`lm()`, 并用函数`anova()`提取其中的方差分析表, 因此`aov(formula)`等价于`anova(lm(formula))`;
- 2) 单因子方差分析还可使用函数`oneway.test()`, 若各水平下数据的方差相等(使用选项`var.equal=TRUE`), 它等同于使用函数`aov()`进行一般的方差分析; 若各水平下数据的方差不相等(使用选项`var.equal=FALSE`), 则它使用Welch(1951)的近似方法进行方差分析;
- 3) 当各水平下的分布未知时, 则采用第七章讲的Kruskal-Wallis秩和检验进行方差分析.

§8.2 双因子方差分析

对于两因素的方差分析, 基本思想和方法与单因素的方差分析相似, 前提条件仍然是要满足独立、正态、方差齐性. 所不同的是在双因素方差分析中, 有时会出现交互作用, 即二因素的不同水平交叉搭配对指标产生影响. 我们先讨论无交互作用的双因素方差分析.

8.2.1 无交互作用的方差分析

设有 AB 两个因素, 因素 A 有 r 个水平 A_1, A_2, \dots, A_r ; 因素 B 有 s 个水平 B_1, B_2, \dots, B_s . 在因素 AB 的每一个水平组合 (A_i, B_j) 下进行一次独立试验得到观察值 $X_{ij}, i = 1, 2, \dots, r, j = 1, 2, \dots, s$, 假定 $X_{ij} \sim N(\mu_{ij}, \sigma^2)$, 且各 X_{ij} 相互独立. 则不考虑交互作用的两因素方差分析模型可表示为

$$\begin{cases} X_{ij} = \mu_i + \alpha_i + \beta_j + \varepsilon_{ij}, i = 1, 2, \dots, r, j = 1, 2, \dots, s, \\ \varepsilon_{ij} \sim N(0, \sigma^2), \text{且各}\varepsilon_{ij}\text{相互独立}, \\ \sum_{i=1}^r \alpha_i = 0, \quad \sum_{j=1}^s \beta_j = 0. \end{cases}$$

其中 $\mu = \frac{1}{rs} \sum_{i=1}^r \sum_{j=1}^s \mu_{ij}$ 为总平均. α_i 为因素 A 的第 i 个水平的效应, β_j 为因素 B 的第 j 个水平的效应.

在给定显著性水平 α 下, 考虑如下假设检验:

$H_{01} : \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$ (因子A对指标影响不显著)

$H_{02} : \beta_1 = \beta_2 = \cdots = \beta_s = 0$ (因子B对指标影响不显著)

类似于单因素方差分析, 先对总离差平方和 SS_T 分解为因素A的效应平方和 SS_A 、因素B的效应平方和 SS_B 及误差平方和 SS_E , 即

$$\begin{aligned}
 SS_T &= \sum_{i=1}^r \sum_{j=1}^s (X_{ij} - \bar{X})^2 \\
 &= \sum_{i=1}^r \sum_{j=1}^s [(x_{yij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}) + (\bar{X}_{i.} - \bar{X}) + (\bar{X}_{.j} - \bar{X})]^2 \\
 &= \sum_{i=1}^r \sum_{j=1}^s (\bar{X}_{i.} - \bar{X})^2 + \sum_{i=1}^r \sum_{j=1}^s (\bar{X}_{.j} - \bar{X})^2 + \sum_{i=1}^r \sum_{j=1}^s (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X})^2 \\
 &= SS_A + SS_B + SS_E
 \end{aligned}$$

其中

$$\begin{aligned}
 \bar{X} &= \frac{1}{rs} \sum_{i=1}^r \sum_{j=1}^s X_{ij} \\
 \bar{X}_{i.} &= \frac{1}{s} \sum_{j=1}^s X_{ij} \quad (i = 1, 2, \cdots, r) \\
 \bar{X}_{.j} &= \frac{1}{r} \sum_{i=1}^r X_{ij} \quad (j = 1, 2, \cdots, s)
 \end{aligned}$$

可以证明:

1) 当 H_{01} 成立时,

$$\frac{SS_A}{\sigma^2} \sim \chi^2(r-1), \quad \frac{SS_E}{\sigma^2} \sim \chi^2((r-1)(s-1)),$$

且 SS_A 与 SS_E 独立, 于是

$$F_A = \frac{SS_A/(r-1)}{SS_E/[(r-1)(s-1)]} \sim F(r-1, (r-1)(s-1)).$$

2) 当 H_{02} 成立时,

$$\frac{SS_B}{\sigma^2} \sim \chi^2(s-1),$$

且 SS_B 与 SS_E 独立, 于是

$$F_B = \frac{SS_B/(s-1)}{SS_E/[(r-1)(s-1)]} \sim F(s-1, (r-1)(s-1)).$$

所以, H_{01} 与 H_{02} 的拒绝域分别为

$$\begin{aligned} C_A &= \{F_A > F_{1-\alpha}(r-1, (r-1)(s-1))\} \\ C_B &= \{F_B > F_{1-\alpha}(s-1, (r-1)(s-1))\}. \end{aligned}$$

在R软件中, 方差分析函数`aov()`既适合于单因素方差分析, 也同样适用于双因素方差分析, 其中方差模型公式为 $x \sim A + B$, 加号表示两个因素具有可加的. 下面用一个例子来说明.

例 8.2.1 原来检验果汁中含铅量有三种方法 A_1 、 A_2 、 A_3 , 现研究出另一种快速检验法 A_4 , 能否用 A_4 代替前三种方法, 需要通过实验考察. 观察的对象是果汁, 不同的果汁当做不同的水平: B_1 为苹果, B_2 为葡萄汁, B_3 为西红柿汁, B_4 为苹果饮料汁, B_5 桔子汁, B_6 菠萝柠檬汁. 现进行双因素交错搭配试验, 即用四种方法同时检验每一种果汁, 其检验结果如表8.4所示. 问因素A(检验方法)和B(果汁品种)对果汁的含铅量是否有显著影响?

表 8.4 果汁含铅比测试实验数据统计

因素	因素B						
A	B_1	B_2	B_3	B_4	B_5	B_6	X_i
A_1	0.05	0.46	0.12	0.16	0.84	1.30	2.93
A_2	0.08	0.38	0.40	0.10	0.92	1.57	3.45
A_3	0.11	0.43	0.05	0.10	0.94	1.10	2.73
A_4	0.11	0.44	0.08	0.03	0.93	1.15	2.74
$X_{.j}$	0.35	1.71	0.65	0.39	3.63	5.12	$X_{..} = 11.85$

解 首先建立数据框:

```
> juice<-data.frame(
  X = c(0.05, 0.46, 0.12, 0.16, 0.84, 1.30, 0.08, 0.38, 0.4,
        0.10, 0.92, 1.57, 0.11, 0.43, 0.05, 0.10, 0.94, 1.10,
        0.11, 0.44, 0.08, 0.03, 0.93, 1.15),
  A = gl(4, 6),
  B = gl(6, 1, 24)
)
```

注: 这里函数`gl()`用来给出因子水平, 其调用格式为

—— `gl()` 的调用格式 ——

```
gl(n, k, length=n*k, labels=1:n, ordered=FALSE)
```

说明: n 是水平数, k 是每一水平上的重复次数, `length`是总观测值数, `ordered`指明各水平是否先排序.

下面作双因素方差分析, **R**程序为:

```
> juice.aov<-aov(X~A+B, data=juice)
> summary(juice.aov)
```

分析结果为

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	3	0.0570	0.0190	1.6287	0.2248
B	5	4.9022	0.9804	83.9755	2.003e-10 ***
Residuals	15	0.1751	0.0117		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

结论: p 值说明果汁品种(因素 B)对含铅量有显著影响, 而没有充分理由说明检验方法(因素 A)对含铅量有显著影响.

最后用函数`bartlett.test()`分别对因素 A 和因素 B 作方差的齐性检验:

```
> bartlett.test(X~A, data=juice) # 对因素A
Bartlett test of homogeneity of variances
```



```
data:  X by A
Bartlett's K-squared = 0.268, df = 3, p-value = 0.966

> bartlett.test(X~B, data=juice) #对因素B
      Bartlett test of homogeneity of variances

data:  X by B
Bartlett's K-squared = 17.4216, df = 5, p-value = 0.003766
```

结论: 对因素A, p 值(0.966)远大于0.05, 接受原假设, 认为因素A的各水平下的数据是等方差的; 对因素B, p 值(0.003766)小于0.05, 拒绝原假设, 即认为因素B不满足方差齐性要求. ■

8.2.2 有交互作用的方差分析

设有两个因素A和B, 因素A有 r 个水平 A_1, A_2, \dots, A_r ; 因素B有 s 个水平 B_1, B_2, \dots, B_s . 在许多情况下, 两因素A与B之间存在着一定程度的交互作用. 为了考察因素间的交互作用, 要求在两个因素的每一水平组合下进行重复试验. 设在每种水平组合 (A_i, B_j) 下重复试验 t 次. 记第 k 次的观测值为 X_{ijk} . 则有交互作用的两因素方差分析模型可表示为

$$\begin{cases} X_{ij} = \mu + \alpha_i + \beta_j + \delta_{ij} + \varepsilon_{ijk}, i = 1, 2, \dots, r, j = 1, 2, \dots, s, k = 1, 2, \dots, t \\ \varepsilon_{ijk} \sim N(0, \sigma^2), \text{且各}\varepsilon_{ijk}\text{相互独立} \\ \sum_{i=1}^r \alpha_i = 0, \quad \sum_{j=1}^s \beta_j = 0, \quad \sum_{i=1}^r \delta_{ij} = \sum_{j=1}^s \delta_{ij} = 0 \end{cases}$$

这里 α_i 为因素A的第 i 个水平的效应, β_j 为因素B的第 j 个水平的效应, δ_{ij} 为 A_i 和 B_j 的交互效应, $\mu = \frac{1}{rs} \sum_{i=1}^r \sum_{j=1}^s \mu_{ij}$.

检验的假设为

$$\begin{aligned} H_{01} &: \alpha_1 = \alpha_2 = \dots = \alpha_r = 0 \quad (\text{因素A对指标X没有影响}) \\ H_{02} &: \beta_1 = \beta_2 = \dots = \beta_s = 0 \quad (\text{因素B对指标X没有影响}) \\ H_{03} &: \delta_{11} = \delta_{12} = \dots = \delta_{rs} = 0 \quad (\text{因素A和B没有联合作用}) \end{aligned}$$

类似于无交互作用的方差分析, 总的离差平方和可分解为

$$\begin{aligned}
 SS_T &= \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t (X_{ijk} - \bar{X})^2 \\
 &= \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t (X_{ijk} - \bar{X}_{ij.})^2 + st \sum_{i=1}^r (\bar{X}_{i..} - \bar{X})^2 \\
 &\quad + rt \sum_{j=1}^s (\bar{X}_{.j.} - \bar{X})^2 + t \sum_{i=1}^r \sum_{j=1}^s (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X})^2 \\
 &= SS_E + SS_A + SS_B + SS_{A \times B}.
 \end{aligned}$$

其中

$$\begin{aligned}
 \bar{X} &= \frac{1}{rst} \sum_{i=1}^r \sum_{j=1}^s \sum_{k=1}^t X_{ijk}, & \bar{X}_{ij.} &= \frac{1}{t} \sum_{k=1}^t X_{ijk}, \\
 \bar{X}_{i..} &= \frac{1}{st} \sum_{j=1}^s \sum_{k=1}^t X_{ijk}, & \bar{X}_{.j.} &= \frac{1}{rt} \sum_{i=1}^r \sum_{k=1}^t X_{ijk}.
 \end{aligned}$$

可以证明,

1) 当 H_{01} 成立时,

$$F_A = \frac{SS_A/(r-1)}{SS_E/[rs(t-1)]} \sim F(r-1, rs(t-1)).$$

2) 当 H_{02} 成立时,

$$F_B = \frac{SS_B/(s-1)}{SS_E/[rs(t-1)]} \sim F(s-1, rs(t-1)).$$

3) 当 H_{03} 成立时,

$$F_{A \times B} = \frac{SS_{A \times B}/[(r-1)(s-1)]}{SS_E/[rs(t-1)]} \sim F((r-1)(s-1), rs(t-1)).$$

R软件中仍用函数`aov()`进行有交互作用的方差分析, 但其中的方差模型格式为 $x \sim A + B + A:B$. 下面用一个例子来全面展示有交互作用方差分析过程.

例 8.2.2 有一个关于检验毒品强弱的试验, 给48只老鼠注射I、II、III三种毒药(因素A), 同时有A、B、C、D 4种治疗方案(因素B), 这样的试验在每一种因素组合下都重复四次测试老鼠的存活时间, 数据如表8.5所示. 试分析毒药和治疗方案以及它们的交互作用对老鼠存活时间有无显著影响.

表 8.5 老鼠存活时间(年)的实验报告

	A		B		C		D	
I	0.31	0.45	0.82	1.10	0.43	0.45	0.45	0.71
	0.46	0.43	0.88	0.72	0.63	0.76	0.66	0.62
II	0.36	0.29	0.92	0.61	0.44	0.35	0.56	1.02
	0.40	0.23	0.49	1.24	0.31	0.40	0.71	0.38
III	0.22	0.21	0.30	0.37	0.23	0.25	0.30	0.36
	0.18	0.23	0.38	0.29	0.24	0.22	0.31	0.33

解 首先以数据框形式输入数据, 并用函数`plot()`作图. 图8.2显示两因素的各水平均有较大差异存在.

```
> rats<-data.frame(
  Time=c(0.31, 0.45, 0.46, 0.43, 0.82, 1.10, 0.88, 0.72, 0.43, 0.45,
        0.63, 0.76, 0.45, 0.71, 0.66, 0.62, 0.38, 0.29, 0.40, 0.23,
        0.92, 0.61, 0.49, 1.24, 0.44, 0.35, 0.31, 0.40, 0.56, 1.02,
        0.71, 0.38, 0.22, 0.21, 0.18, 0.23, 0.30, 0.37, 0.38, 0.29,
        0.23, 0.25, 0.24, 0.22, 0.30, 0.36, 0.31, 0.33),
  Toxicant=gl(3, 16, 48, labels = c("I", "II", "III")),
  Cure=gl(4, 4, 48, labels = c("A", "B", "C", "D"))
)
> op<-par(mfrow=c(1, 2))
> plot(Time~Toxicant+Cure, data=rats)
```

下面再用函数`interaction.plot()`作出交互效应图, 以考查因素之间交互作用是否存在, **R**程序为

```
> with(rats,
```

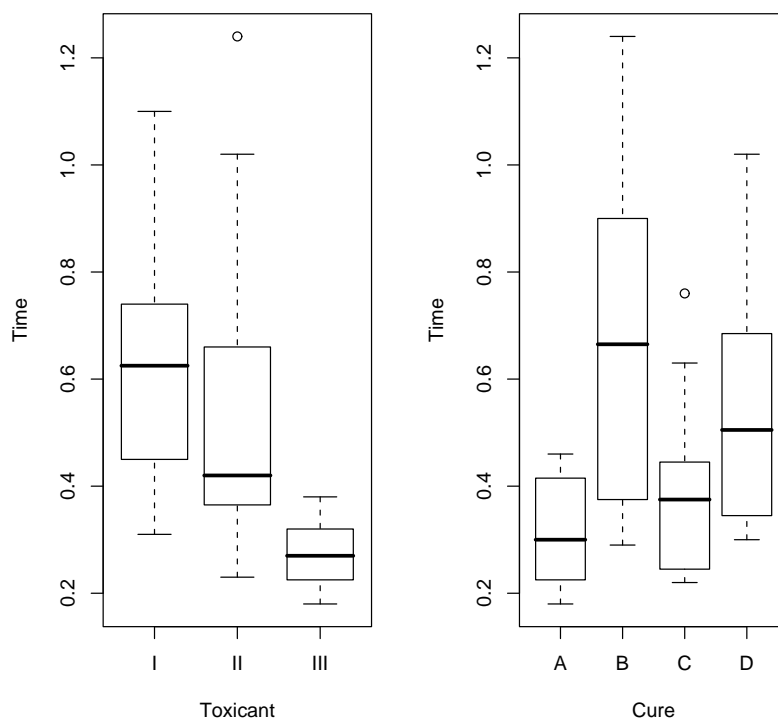


图 8.2 毒药和治疗方案两因素的各自效应分析

```
interaction.plot(Toxicant, Cure, Time, trace.label="Cure")
> with(rats,
  interaction.plot(Cure, Toxicant, Time, trace.label="Toxicant"))
```

输出结果如图8.3(a)和图8.3(b). 两图中的曲线并没有明显的相交情况出现, 因此我们初步认为两个因素没有交互作用.

尽管如此, 由于实验误差的存在, 我们用方差分析函数`aov()`对此进行确认, 其中方差模型格式为 $x A * B$, 或 $A + B + A : B$, 表示不仅考虑因素 A 、 B 各自的效应, 还考虑两者的交互效应. 若仅考虑 A 与 B 的交互效应则方差模型格式为 $A : B$.

由R程序

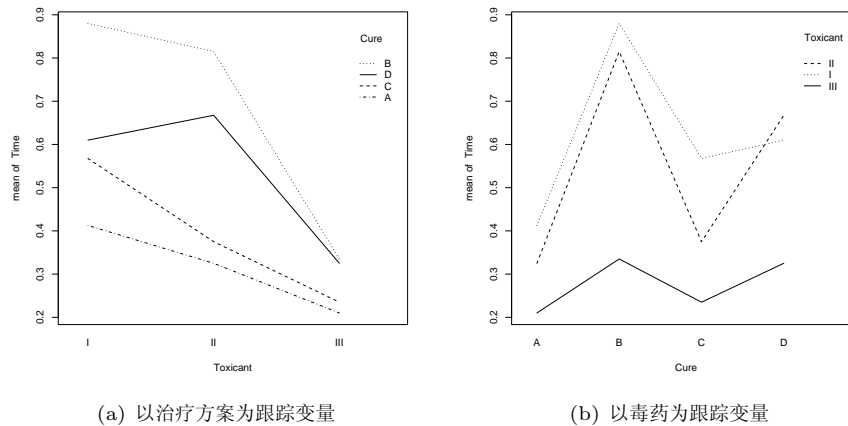


图 8.3 交互效应图

```
> rats.aov<-aov(Time~Toxicant*Cure, data=rats)
> summary(rats.aov)
```

得到检验结果为

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Toxicant	2	1.03563	0.51781	23.2254	3.326e-07 ***
Cure	3	0.91462	0.30487	13.6745	4.132e-06 ***
Toxicant:Cure	6	0.24782	0.04130	1.8526	0.1163
Residuals	36	0.80262	0.02230		

 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

根据 p 值知, 因素Toxicant和Cure对Time的影响是高度显著的, 而交互作用对Time的影响却是不显著的.

再进一步使用前面的Bartlett和Levene两种方法检验因素Toxicant和Cure下的数据是否满足方差齐性的要求, R程序如下.

```
> library(car)
> levene.test(rats$Time, rats$Toxicant)
> levene.test(rats$Time, rats$Cure)
> bartlett.test(Time~Toxicant, data=rats)
```

```
> bartlett.test(Time~Cure, data=rats)
```

结果显示从略, 其中各 p 值均小于0.05表明在0.05显著性水平下两因素下的方差不能满足齐性的要求, 这与图8.2是一致的. ■

§8.3 协方差分析

前面两节介绍的方差分析方法中两组或多组均值间比较的假设检验, 其处理因素一般是可以控制的. 但在实际工作中, 有时有些因素无法加以控制, 如何在比较两组或多组均数间差别的同时扣除或均衡这些不可控因素的影响, 可考虑采用协方差分析的方法.

协方差分析(Analysis of Covariance, 简称ancova)是将线性回归分析与方差分析结合起来的一种统计分析方法. 其基本思想就是: 将一些对响应变量 Y 有影响的变量(指未知或难以控制的因素)看作协变量(covariate), 建立响应变量 Y 随协变量 X 变化的线性回归关系, 并利用这种回归关系把 X 值化为相等后再对各处理组 Y 的修正均值(adjusted means)间差别进行假设检验, 其实质就是从 Y 的总的平方和中扣除 X 对 Y 的回归平方和, 对残差平方和作进一步分解后再进行方差分析, 以更好地评价这种处理的效应.

可见, 对于一个协方差分析模型, 方差分析是主要的, 我们的基本目的是作方差分析, 而回归分析仅仅是因为回归变量(协变量)不能完全控制而引入的. 下面讨论最简单的情形: 一个协变量、单因素的协方差分析.

设试验只有一个因素 A 在变化, A 有 r 个水平 A_1, A_2, \dots, A_r , 与之有关的仅有一个协变量 X , 在水平 A_i 下进行 n_i 次独立观测, 得到 n 对观测数据 $(X_{ij}, Y_{ij}), i = 1, 2, \dots, r; j = 1, 2, \dots, n_i$, 则协方差模型可用线性模型表示为

$$\begin{cases} Y_{ij} = \mu + \alpha_i + \beta(X_{ij} - \bar{X}_{..}) + \varepsilon_{ij}, i = 1, 2, \dots, r, j = 1, 2, \dots, n_i \\ \varepsilon_{ij} \sim N(0, \sigma^2), \text{且各}\varepsilon_{ij}\text{相互独立} \\ \sum_{i=1}^r n_i \alpha_i = 0, \quad \beta \neq 0 \end{cases}$$

其中 μ 为总平均, α_i 为第 i 个水平的效应, β 是 Y 对 X 的线性回归函数, ε_{ij} 为随机误差, $\bar{X}_{..}$ 为 X_{ij} 的总平均数.

给定显著水平 α , 考虑假设检验

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0, \longleftrightarrow H_1: \alpha_1, \alpha_2, \cdots, \alpha_r \text{不全相等}$$

令

$$SS_T(y) = \sum_{i=1}^r \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^r \sum_{j=1}^{n_i} Y_{ij}^2 - n\bar{Y}_{..}^2 \quad (y \text{ 的总离均差平方和})$$

$$SS_A(y) = \sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2 = \sum_{i=1}^r n_i \bar{Y}_{i.}^2 - n\bar{Y}_{..}^2 \quad (y \text{ 的组间平方和})$$

$$SS_E(y) = \sum_{i=1}^r \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 = SS_T(y) - SS_A(y) \quad (y \text{ 的组内平方和})$$

$$SS_T(x) = \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})^2 = \sum_{i=1}^r \sum_{j=1}^{n_i} X_{ij}^2 - n\bar{X}_{..}^2 \quad (x \text{ 的总离均差平方和})$$

$$SS_A(x) = \sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{X}_{i.} - \bar{X}_{..})^2 = \sum_{i=1}^r n_i \bar{X}_{i.}^2 - n\bar{X}_{..}^2 \quad (x \text{ 的组间平方和})$$

$$SS_E(x) = \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2 = SS_T(x) - SS_A(x) \quad (x \text{ 的组内平方和})$$

$$SP_T = \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})(Y_{ij} - \bar{Y}_{..}) = \sum_{i=1}^r \sum_{j=1}^{n_i} X_{ij}Y_{ij} - n\bar{X}_{..}\bar{Y}_{..} \quad (x \text{ 与 } y \text{ 的总离均差乘积和})$$

$$SP_A = \sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{X}_{i.} - \bar{X}_{..})(\bar{Y}_{i.} - \bar{Y}_{..}) = \sum_{i=1}^r n_i \bar{X}_{i.}\bar{Y}_{i.} - n\bar{X}_{..}\bar{Y}_{..} \quad (x \text{ 与 } y \text{ 的组间乘积和})$$

$$SP_E = \sum_{i=1}^r \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})(Y_{ij} - \bar{Y}_{i.}) = SP_T - SP_A \quad (x \text{ 与 } y \text{ 的组内乘积和})$$

其中

$$\bar{X}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \quad \bar{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij},$$

$$\begin{aligned}\bar{X}_{..} &= \frac{1}{r} \sum_{i=1}^r \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} = \frac{1}{r} \sum_{i=1}^r \bar{X}_{i.}, \\ \bar{Y}_{..} &= \frac{1}{r} \sum_{i=1}^r \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{r} \sum_{i=1}^r \bar{Y}_{i.}.\end{aligned}$$

由此得参数 μ 、 α_i 和 β 的估计为

$$\hat{\mu} = \bar{Y}_{..}, \quad \hat{\beta} = b^* = \frac{SP_E}{SS_E(x)}, \quad \hat{\alpha}_i = \bar{Y}_{i.} - \bar{Y}_{..} - b^*(\bar{X}_{i.} - \bar{X}_{..})$$

其中 $b^*(\bar{X}_{i.} - \bar{X}_{..})$ 反映了因线性回归系数显著时对数据知矫正. 这时矫正后的组内平方和为

$$SS_E = SS_E(y) - b^* SP_E = SS_E(y) - \frac{SP_E^2}{SS_E(x)},$$

其自由度为 $df = n - r - 1$, 且 $\frac{SS_E}{\sigma^2} \sim \chi^2(n - r - 1)$. 矫正后总平方和为

$$SS_T = SS_T(y) - \frac{SP_T^2}{SS_T(x)},$$

矫正后的组间平方和为

$$SS_A = SS_T - SS_E,$$

其自由度为 $df = r - 1$, 且 $\frac{SS_A}{\sigma^2} \sim \chi^2(r - 1)$. 而在 H_0 成立时, SS_A 与 SS_E 独立, 从而

$$F = \frac{\frac{SS_A}{r-1}}{\frac{SS_E}{n-r-1}} \sim F(r-1, n-r-1).$$

因此, 若 $F > F_{1-\alpha}(r-1, n-r-1)$, 则拒绝 H_0 , 即认为各水平效应显著不同. 反之“接受”原假设.

R中**HH**程序包中的函数`ancova()`提供了方差分析的计算, 其调用格式为

—— `ancova()` 的调用格式 ——

```
ancova(formula, data.in = sys.parent(),
       x, groups)
```

说明:`formula`是协方差分析的公式, `data.in`是数据框, `x`为协方差分析中的

协变量, 在作图时若formula中没有x则需要指出, groups为因子, 在作图时若formula的条件项中没有groups则需要指出, 其它参见在线帮助.

例 8.3.1 为研究A、B、C三种饲料对猪的催肥效果, 用每种饲料喂养8头猪一段时间, 测得每头猪的初始重量(X)和增重(Y), 数据见表8.6. 试分析三种饲料对猪的催肥效果是否相同?

表 8.6 三种饲料喂养猪的初始重量与增重

	饲料A		B饲料		C饲料	
	X_1	Y_1	X_2	Y_2	X_3	Y_3
1	15	85	17	97	22	89
2	13	83	16	90	24	91
3	11	65	18	100	20	83
4	12	76	18	95	23	95
5	12	80	21	103	25	100
6	16	91	22	106	27	102
7	14	84	19	99	30	105
8	17	90	18	94	32	110

解 饲料是人为可以控制的定性因素, 猪的初始重量是难以控制的定量因子, 为协变量X; 实验的观察指标是猪的增量, 为响应变量Y. 各组的增重由于受猪的原始体重影响, 不能直接进行方差分析, 需进行协方差分析. **R**程序及结果如下:

- 建立数据集

```
feed<-rep(c("A","B","C"),each=8)
Weight_Initial <- c(15,13,11,12,12,16,14,17,17,16,
                    18,18,21,22,19,18,22,24,20,23,
                    25,27,30,32)
Weight_Increment <-c(85,83,65,76,80,91,84,90,97,90,
                     100,95,103,106,99,94,89,91,83,
                     95,100,102,105,110)
```

```
data_feed<-data.frame(feed,Weight_Initial,Weight_Increment)
```

- 若认为在三种不同饮料喂养下, 猪的初始体重不同, 但增长速度相同, 这时使用命令

```
> ancova(Weight_Increment ~ Weight_Initial+feed ,
        data=data_feed)
```

得到图形8.4. 由函数anov()提取协方差分析表得

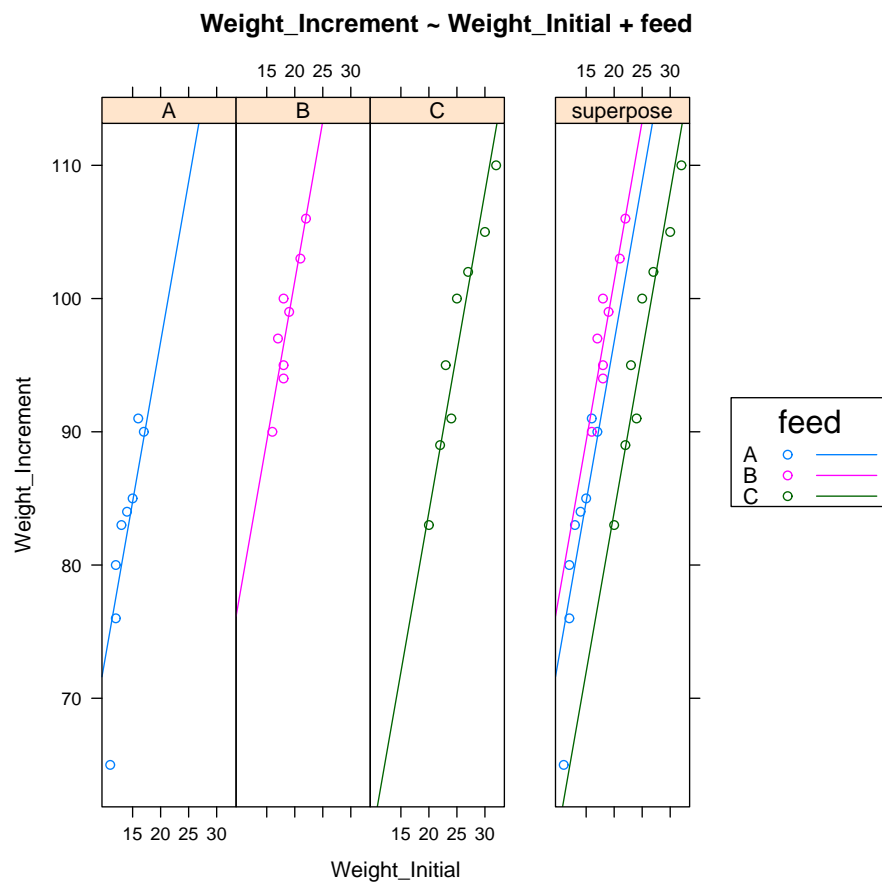


图 8.4 增长速度相同下的回归

Analysis of Variance Table

```

Response: Weight_Increment
              Df  Sum Sq Mean Sq F value    Pr(>F)
Weight_Initial  1 1621.12 1621.12 142.445 1.496e-10 ***
feed            2  707.22  353.61  31.071 7.322e-07 ***
Residuals      20  227.61   11.38
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

```

可见猪的初始体重和增长速度对猪的增重都有显著差异。

- 若认为在三种不同饮料喂养下, 猪的初始体重和增长速度都不同, 这时使用命令

```

> ancova(Weight_Increment ~ Weight_Initial*feed ,
          data=data_feed)

```

得到图形8.5.

从两个图的比较及初始体重对增长的检验发现, 三种饲料对猪的催肥效果相同, 猪的初始体重对其影响不大. ■

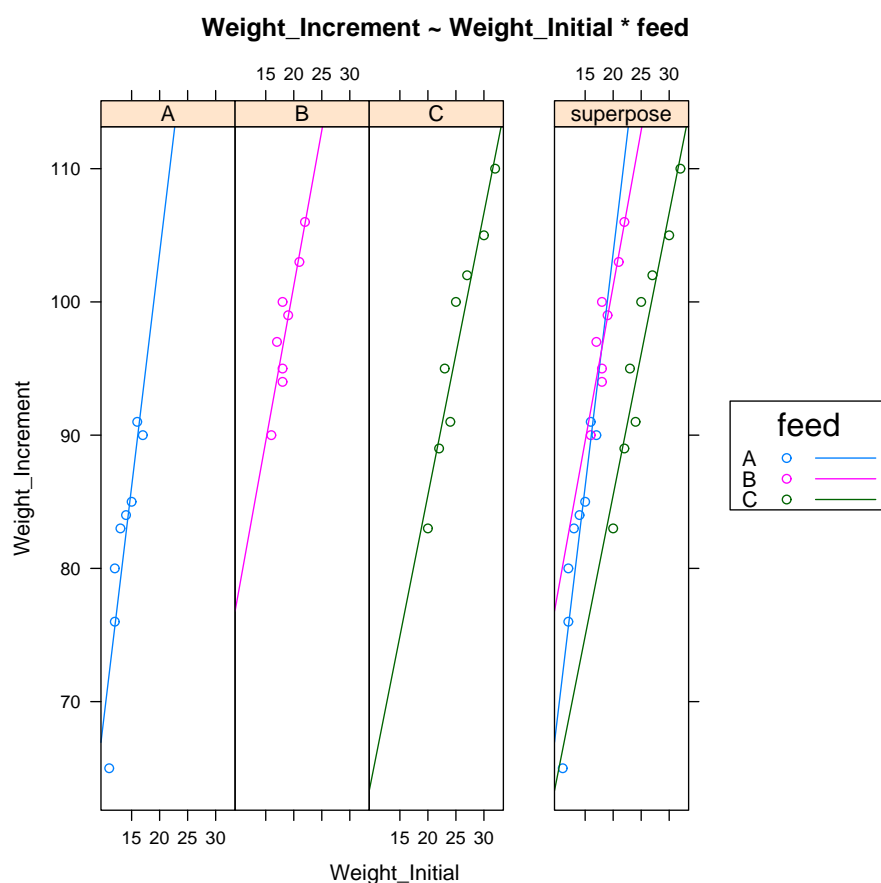


图 8.5 增长速度不同下的回归

第八章习题

8.1 有4个不同的实验室制作同一型号的纸张, 为比较各实验室生产纸张的光滑度, 测量了每个实验室生产的8种张纸, 得其光滑度如表8.7所示. 假设上述数据服从方差分析模型. 试在显著性水平 $\alpha = 0.05$ 下, 检验各个实验室生产的纸张的光滑度是否有显著差异.

8.2 在对比研究中观察正常人、萎缩性胃炎和胃癌三个不同群体(用TYPE=A, B和C表示), 记录的资料见表8.8, 试对该组数据作方差分析.

- 1) 检验三个群体中CEA含量的分布是否为正态分布, 方差是否相等($\alpha =$

表 8.7 四个实验室生产纸张的光滑度

实验室	纸张光滑度							
A_1	38.7	41.5	43.8	44.5	45.5	46.0	47.7	58.0
A_2	39.2	39.3	39.7	41.4	41.8	42.9	43.3	45.8
A_3	34.0	35.0	39.0	40.0	43.0	43.0	44.0	45.0
A_4	34.0	34.8	34.8	35.4	37.2	37.8	41.2	42.8

0.01)?

2) 试用方差分析(ANOVA)过程比较这三个群体CEA含量有无显著差异? 若有显著差异, 请指出哪些群体间CEA的平均含量有显著差异? ($\alpha = 0.05$)

8.3 在饲料对样鸡增肥的研究中, 某研究所提出三种饲料配方: A_1 是以鱼粉为主的饲料, A_2 是以槐树粉为主的饲料, A_3 是以苜蓿粉为主的饲料, 为比较三种饲料的效果, 特选30只雏鸡随机均分为三组, 每组各喂一种饲料, 60天后观察它们的重量, 试验结果如表8.9所示. 度在显著性水平 $\alpha = 0.05$ 下进行方差分析, 可以得到哪些结果?

8.4 为考察对纤维弹性测量的误差, 现对四个工厂(A_1, A_2, A_3, A_4)生产的同一批原料进行测量, 每厂各找四个检验员(B_1, B_2, B_3, B_4)轮流使用各厂设备进行重复测量, 试验数据如表8.10所示. 请问因素A与B的影响是否显著($\alpha = 0.05$)?

8.5 水稻试验问题: 考察的因子有水稻品种A和施肥量B; 考察的指标为水稻的产量Y. 设因子A有三个水平: A_1 (窄叶青), A_2 (珍珠矮) 和 A_3 (江二矮); 因子B有四个水平: B_1 (无肥), B_2 (低肥), B_3 (中肥)和 B_4 (高肥). 对这12种搭配的每一种, 在两块试验田上做实验. 每块试验田分为12块面积相同的小田, 随机地安排12种搭配条件进行试验. 得数据如下表8.11所示. 试分析水稻试验数据, 并回答以下问题:

- 1) 不同稻种的产量是否有显著的差别?哪种稻种更好些?
- 2) 不同的施肥量对产量是否有明显的影响? 最适合的施肥量是多少?
- 3) 稻种和施肥量对产量的影响哪个更大些?
- 4) 稻种和施肥量有无交互作用?

表 8.8 胃液癌胚抗原(CEA)含量X(mg/ml)

正常人 (A)	20.4	30.2	210.4	365.0	56.8	37.8
	265.3	175.0	169.8	356.4	254.0	262.3
	170.5	360.0	78.4	86.4	128.0	24.1
	28.5	108.5	472.5	158.6	238.7	253.6
	57.0	189.6	59.3	259.3	380.2	210.5
	64.6	87.3				
萎缩性 胃炎 (B)	281.0	377.1	230.0	537.9	248.7	571.4
	766.2	495.0	87.3	389.8	423.9	577.3
	66.8	521.3	327.8	421.4	149.7	47.5
	425.7	270.8	378.5	228.0	538.7	245.6
	584.1	64.8	485.6	110.8	398.7	452.6
	587.7	86.8	532.1	311.6	442.2	
胃癌 (C)	480.0	488.9	350.7	652.8	1400.0	850.0
	725.6	590.0	765.0	1200.0	231.2	485.3
	600.0	1380.0	438.5	652.4	432.8	296.1
	464.8	608.4	688.5	630.5	750.0	815.0
	664.0	348.6	550.0	640.0		

表 8.9 鸡饲料试验数据

饲料	鸡重/g									
A_1	1073	1058	1071	1037	1066	1026	1053	1049	1065	1051
A_2	1061	1058	1038	1042	1020	1045	1044	1061	1034	1049
A_3	1084	1069	1106	1078	1075	1090	1079	1094	1111	1092

5) 使产量达到最高的生产条件是什么?

8.6 用3种压力(B_1 、 B_2 、 B_3)和四种温度(A_1 、 A_2 、 A_3 、 A_4)组成的集中试验

表 8.10 纤维弹性数据

检验员	A ₁	A ₂	A ₃	A ₄
B ₁	71.73	73.75	76.73	71.73
B ₂	72.73	76.74	79.77	73.72
B ₃	75.73	78.77	74.75	70.71
B ₄	77.75	76.74	74.73	69.69

表 8.11 水稻试验数据

	B ₁		B ₂		B ₃		B ₄	
A ₁	19.3	19.2	24.0	27.3	26.0	28.5	27.8	28.5
A ₂	21.7	22.6	27.5	30.3	29.0	28.7	30.2	29.8
A ₃	20.0	20.1	24.2	27.3	24.5	27.1	28.1	27.7

方案, 得到产品得率资料如表8.12所示, 试分析压力和温度以及它们的交互作用对产品得率有无显著影响($\alpha = 0.05$).

表 8.12 实验数据及计算表

温度A	压力B								
	B ₁			B ₂			B ₃		
A ₁	52	43	39	41	47	53	49	38	42
A ₂	48	37	39	50	41	30	36	48	47
A ₃	34	42	38	36	39	44	37	40	32
A ₄	45	58	42	44	46	60	43	56	41

8.7 为了提高化工厂的产品质量, 需要寻求最优反应速度与压力的搭配, 为此选择如下水平:

A: 反应速度(m/s) 60 70 80,

B: 反应应力(kg) 2 2.5 3

在每个 (A_i, B_j) 条件下做2次试验, 其产量如表8.13所示.

表 8.13 试验数据

	A_1		A_2		A_3	
B_1	4.6	4.3	6.1	6.5	6.8	6.4
B_2	6.3	6.7	3.4	3.8	4.0	3.8
B_3	4.7	4.3	3.9	3.5	6.5	7.0

(1)对数据作方差分析(应考虑交互作用);

(2)对 (A_i, B_j) 条件下平均产量作多重比较.

8.8 在庆大霉素三种不同水平下, 即对照组(无), 30ug/ml和300ug/ml作兔子结肠器官培养液中胸腺嘧啶核苷的吸收分析. 对每个试验, 可得到不同浓度的胸腺嘧啶核苷 X 的含量和DNA的含成量 Y , 数据如表8.14所示. 如果抗生素有效, DNA(Y)合成率会降低. 试作协方差分析.

表 8.14 在兔结肠器官培养液中胸腺嘧啶核苷的吸收分析

对照		30ug/ml		300ug/ml	
X_1	Y_1	X_2	Y_2	X_3	Y_3
1.40	0	1.6	0	2.2	0
1.5	3	2.0	3	2.3	3
1.8	5	2.3	5	3.0	10
2.2	10	2.9	10	3.2	5
3.4	2.	4.5	20	4.5	20
3.6	25	5.1	25	5.9	30
4.6	30	6.0	25	7.0	30

8.9 已知出生体重随种族的不同而不同. 白种人婴儿的出生体重比其他种族的重. 出生体重也随孕期的增长而增加, 足月(40周)的婴儿通常比不足月(小

于40周)的重. 一般来说, 当比较不同种族婴儿的出生体重时必须对孕期长短进行校正. 表8.15是出生体重孕期长短的数据. 试作协方差分析.

表 8.15 按母亲种族分类的出生体重的孕期

白人		黑人		西班牙人		亚洲人	
孕期 (天数)	出生体重 (盎司)	孕期 (天数)	出生体重 (盎司)	孕期 (天数)	出生体重 (盎司)	孕期 (天数)	出生体重 (盎司)
X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
260	130	260	115	262	113	260	111
275	135	263	118	264	115	271	174
278	138	270	120	270	120	274	117
280	142	278	125	275	121	279	118
282	146	281	128	280	127	281	120
288	149	285	132	284	132	283	122

第九章 回归分析与相关分析

本章概要

- ◇ 相关性及其度量
- ◇ 一元线性回归分析
- ◇ 多元线性回归分析
- ◇ 回归诊断
- ◇ logistic回归

相关分析和回归分析是研究变量间相互关系,测定它们联系的紧密程度.揭示其变化的具体形式和规律性的统计方法,是构造各种经济模型、进行结构分析、政策评价、预测和控制的重要工具.

§9.1 相关性及其度量

9.1.1 相关性概念

变量之间相互关系大致可分为两种类型,即函数关系和相关关系.函数关系是指变量之间存在的相互依存关系,它们之间的关系可以用某一方程(函数) $y = f(x)$ 表达出来;相关关系是指两个变量的数值变化存在不完全确定的依存关系,它们之间的数值不能用方程表示出来,但可用某种相关性度量来刻画.相关关系是相关分析的研究对象,而函数关系则是回归分析的研究对象.

相关的种类繁多,按照不同的标准可有不同的划分.按照相关程度的不同,可分为完全相关、不完全相关、不相关;按照相关方向的不同,可分为正相关和负相关;按照相关形式的不同,又可分为线性相关和非线性相关;按涉及变量的

多少可分为一元相关和多元相关; 按影响因素的不同, 可分为单相关和复相关.

在进行相关分析和回归分析之前, 可先通过不同变量之间的散点图直观地了解它们之间的关系和相关程度. 常见的是一些连续变量间的散点图, 若图中数据点分布在一条直线(曲线)附近, 表明可用直线(曲线)近似地描述变量间的关系. 若有多个变量, 常制作多幅两两变量间的散点图来考察变量间的关系.

R中使用函数`plot()`可以方便地画出两个样本的散点图, 从而直观地了解对应随机变量之间的相关关系和相关程度.

例 9.1.1 某医生测定了10名孕妇的15-17周及分娩时脐带血TSH (Mu/L)水平. 试绘制脐带血和母血的散点图.

表 9.1 10名孕妇的15-17周及分娩时脐带血TSH(Mu/L)

母血TSH(X)	1.21	1.30	1.39	1.42	1.47	1.56	1.68	1.72	1.98	2.10
脐带血(Y)	3.90	4.50	4.20	4.83	4.16	4.93	4.32	4.99	4.70	5.20

解 **R**程序如下:

```
> x<-c(1.21, 1.30, 1.39, 1.42, 1.47, 1.56, 1.68, 1.72, 1.98, 2.10)
> y<-c(3.90, 4.50, 4.20, 4.83, 4.16, 4.93, 4.32, 4.99, 4.70, 5.20)
> level <- data.frame(x,y)
> plot(level)
```

运行结果如图9.1. 从图上可以直观看出, 数据点分布相对较为分散, 但观察所有点的分布趋势, 又可能存在某种递增的趋向, 所以可推测X和Y之间有某种正相关关系. ■

9.1.2 相关分析

散点图是一种最为有效最为简单的相关性分析工具. 若通过散点图可以基本明确它们之间存在直线关系, 则可通过线性回归进一步确定它们之间的函数关系(见§9.2), 它们之间的相关程度可以用**Person**相关系数来刻画, 因此**Person**相关系数实际上反映了变量间的线性相关程度的大小. 除此之外, 还有**Spearman**秩相关系数和**Kendall**相关系数.

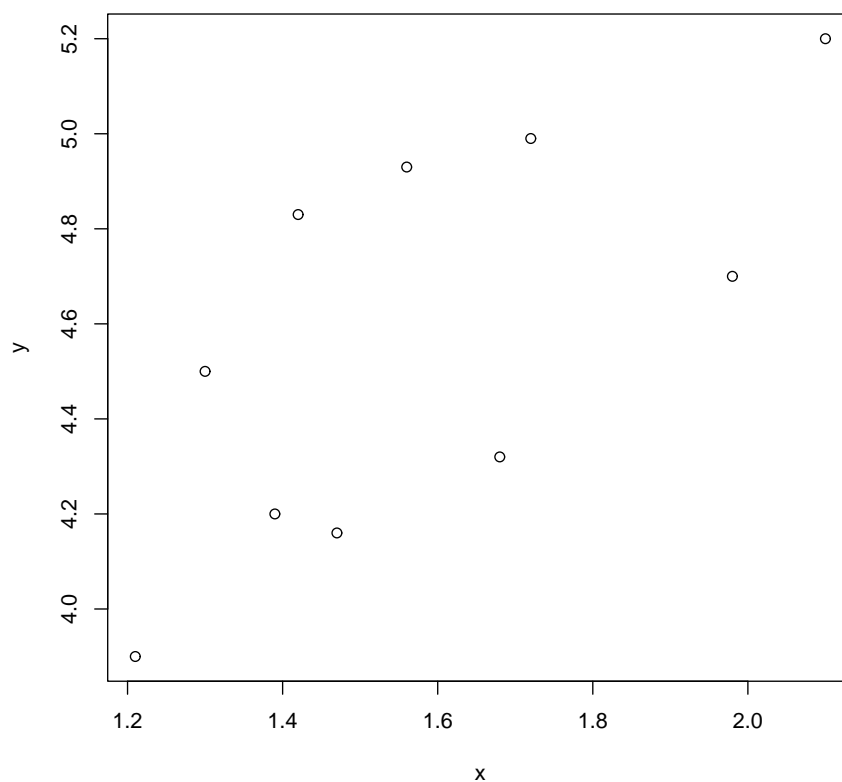


图 9.1 脐带血与母血TSH数据点的散点图

设两个随机变量 X 与 Y 的观测值为 $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$, 则它们之间的(样本)相关系数为:

$$\gamma_{(X,Y)} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

可以证明, 当样本个数 n 充分大时, 样本相关系数可以作为总体 X 和 Y 的相关系数

$$\rho_{(X,Y)} = \frac{E(X - E(Y))(Y - E(Y))}{\sqrt{Var(X)Var(Y)}}$$

的估计. 因此 $|\gamma| \leq 1$. 当 $|\gamma| \rightarrow 1$ 时, 表明两变量的数据有较强线性关系; 当 $|\gamma| \rightarrow 0$ 时, 表明两变量的数据间几乎无线性关系, $\gamma \geq 0 (\leq 0)$ 表示正(负)相关, 表示随 x 的递增(减), y 的值大体上会递增(减).

进一步, 若 (X, Y) 服从二元正态分布, 则

$$T = \frac{\gamma_{xy}\sqrt{n-2}}{\sqrt{1-\gamma_{xy}^2}} \sim t(n-2).$$

由此可以对 X 和 Y 进行Pearson相关性检验: 若 $T > t_{1-\alpha}(n-2)$, 则认为 X 和 Y 的观测值之间存在显著的(线性)相关性. 此外, 还可根据Spearman秩相关系数和Kendall相关系数进行相应的Spearman秩检验和Kendall检验. 这里只介绍R中的函数, 有关检验原理请参见数理统计教材.

在R软件中, `cor.test()`提供了上述三种检验方法, 其调用格式为

—— `cor.test()` 的调用格式-1 ——

```
cor.test(x, y,
         alternative=c("two.sided", "less", "greater"),
         method=c("pearson", "kendall", "spearman"),
         exact=NULL, conf.level=0.95...)
```

说明: x, y 是长度相同的向量; `alternative`是备择假设, 默认值为“two.sided”; `method`是选择检验方法, 默认值为Pearson检验; `coef.level`是置信水平, 默认值为0.95.

`cor.test()`函数还有另外一种调用格式

—— `cor.test()` 的调用格式-2 ——

```
cor.test(formula, data, subset, na.action, ...)
```

说明: `formula`是公式, 形如‘ $u+v$ ’, ‘ u ’, ‘ v ’, 它们必须是具有相同长度的数值向量; `data`是数据框; `subset`是可选择向量, 表示观察值的子集.

例 9.1.2 对例9.1.1中的两组数据进行相关性检验.

解 R程序如下:

```
> attach(level)
> cor.test(x, y)
```

运行结果为:

```
Pearson's product-moment correlation
data:  x and y t = 2.6284,  df = 8,
p-value = 0.03025 alternative hypothesis:
true correlation is not equal to 0 95 percent confidence interval:
 0.0894336 0.9172270
sample estimates:
      cor
0.6807283
```

结论: 因为 p 值 $=0.03025 \leq 0.05$, 故拒绝原假设, 从而认为变量 x 与 y 相关. ■

§9.2 一元线性回归分析

相关分析只能得出两个变量之间是否相关, 但却不能回答在两个变量之间存在相关关系时, 它们之间是如何联系的, 即无法找出刻画它们之间因果关系的函数关系. 回归分析就可以解决这一问题, 先从一元线性回归讲起.

9.2.1 数学模型

设变量 x 和 y 之间存在一定的相关关系, 回归分析方法即找出 Y 的值是如何随 X 的值的变化的规律, 我们称 Y 为因变量(或响应变量), X 为自变量(或解释变量), 现通过例子说明如何来确定 Y 与 X 之间的关系.

例 9.2.1 有10个同类企业的生产性固定资产价值(X)和工业总产值(Y)资料如下(见表9.2):

为了直观起见, 可画一张“散点图”, 以 x 为横坐标, y 为纵坐标, 每一数据对 (x_i, y_i) 为 X - Y 坐标中的一个点, $i = 1, 2, \dots, 10$, 如下图9.2所示. 相应的命令为

```
> x <- c(318, 910, 200, 409, 425, 502, 314, 1210, 1022, 1225)
> y <- c(524, 1019, 638, 815, 913, 928, 605, 1516, 1219, 1624)
> plot(x, y)
```

从图上发现, 10个点基本在一条直线附近, 从而可以认为 Y 与 X 的关系基本上

表 9.2 企业固定资产价值和工业总产值

企业编号	生产性固定资产价值 (万元)	工业总产值 (万元)
1	318	524
2	910	1019
3	200	638
4	409	815
5	425	913
6	502	928
7	314	605
8	1210	1516
9	1022	1219
10	1225	1624
合计	6525	9801

是线性的, 而这些点与直线的偏离是由其它一切不确定因素造成的, 为此可作如下假定

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad (9-2.1)$$

其中 $Y = \beta_0 + \beta_1 X$ 表示 Y 随 X 变化而线性变化的部分; ε 是随机误差, 它是其它一切不确定因素影响的总和, 其值不可观测, 通常假定 $\varepsilon \sim N(0, \sigma^2)$. 称函数 $f(X) = \beta_0 + \beta_1 X$ 为一元线性回归函数, β_0 为回归常数, β_1 为回归系数, 统称回归参数. 称 X 为回归自变量(或回归因子), Y 为回归因变量(或响应变量).

若 $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ 是 (X, Y) 的一组观测值(样本), 则一元线性回归模型可表示为

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, 2, \dots, n, \quad (9-2.2)$$

其中 $E(\varepsilon_i) = 0, \text{Var}(\varepsilon_i) = \sigma^2, i = 1, 2, \dots, n$.

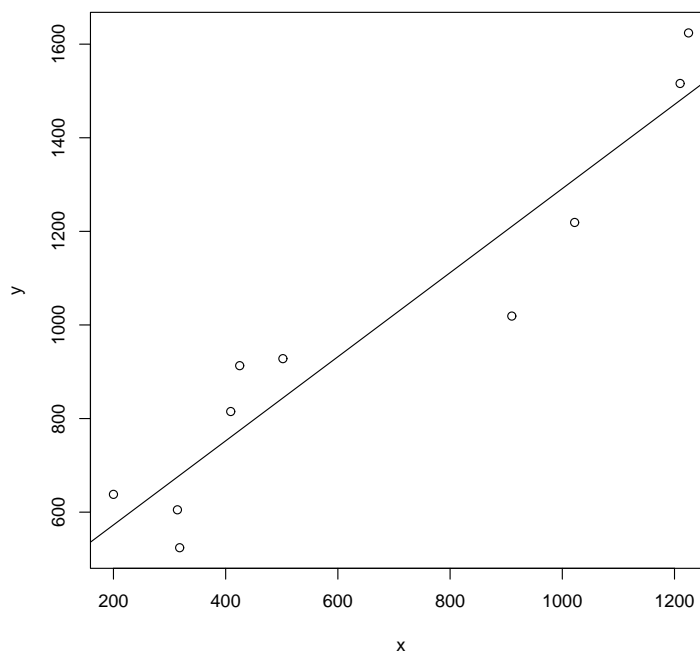


图 9.2 生产性固定资产价值与工业总产值的散点图

9.2.2 估计与检验

β_0, β_1 的估计

求出未知参数 β_0, β_1 的估计 $\hat{\beta}_0, \hat{\beta}_1$ 的一种直观想法就是要求图 9.2 中的点 (X_i, Y_i) 与直线上的点 \hat{X}_i, \hat{Y}_i 的偏离越小越好, 这里 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ 称为回归值或拟合值. 令

$$Q(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2 \quad (9-2.3)$$

则 β_0, β_1 的最小二乘估计就是使 $Q(\beta_0, \beta_1)$ 取得最小值时的 $\hat{\beta}_0, \hat{\beta}_1$.

用微分法可得

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{S_{xy}}{S_{xx}} \quad (9-2.4)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \quad (9-2.5)$$

其中

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad S_{xx} = \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad S_{xy} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

即(一元)回归方程为 $Y = \hat{\beta}_0 + \hat{\beta}_1 X$.

通常取 $\hat{\sigma}^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 / (n-2)$ 为参数 σ^2 的估计量(也称为 σ^2 的最小二乘估计). 可以证明 $E(\hat{\sigma}^2) = \sigma^2$.

回归方程的显著性检验

从回归参数的估计公式(9-2.4)可知, 在计算过程中并不一定要知道 Y 与 X 是否有线性相关的关系, 但如果不存在这种关系, 那么求得的回归方程毫无意义. 因此, 需要对回归方程进行显著性检验. 对于一元线性回归模型, 它等价于回归系数 β_1 的显著性检验.

对于检验问题

$$H_0 : \beta_1 = 0 \leftrightarrow H_1 : \beta_1 \neq 0$$

通常采用三种(等价)的检验方法:

(1) t 检验法. 当 H_0 成立时, 统计量

$$T = \frac{\hat{\beta}_1}{\text{Sd}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 \sqrt{S_{xx}}}{\hat{\sigma}} \sim t(n-2) \quad (9-2.6)$$

对给定的显著性水平 α , 检验的拒绝域为

$$C = \{|T| \geq t_{1-\alpha/2}(n-2)\}$$

(2) F检验法. 当 H_0 成立时, 统计量

$$F = \frac{\hat{\beta}_1 S_{xx}}{\hat{\sigma}_2^2} \sim F(1, n-2) \quad (9-2.7)$$

对于给定的显著性水平 α , 检验的拒绝域为

$$C = \{F = F_{1-\alpha}(1, n-2)\}.$$

(3) 相关系数检验法. 记样本相关系数可表示为 $\gamma_{(X,Y)} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$. 对于给定的显著性水平 α , 检验的拒绝域为

$$C = \{|\gamma_{(X,Y)}| > \gamma_{1-\alpha}(n-2)\}. \quad (9-2.8)$$

上述三种检验中, 当拒绝 H_0 时, 就认为线性回归方程是显著的.

β_0, β_1 的区间估计

由 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 的统计性质知

$$T_i = \frac{\hat{\beta}_i - \beta_i}{\text{Sd}(\beta_i)} \sim t(n-2), \quad i = 0, 1 \quad (9-2.9)$$

因此, 对给定的置信水平 $1 - \alpha$, 由

$$P \left\{ \left| \frac{\hat{\beta}_i - \beta_i}{\text{Sd}(\beta_i)} \right| \leq t_{\frac{\alpha}{2}}(n-2) \right\} = 1 - \alpha, \quad i = 0, 1 \quad (9-2.10)$$

得 $\beta_i (i = 0, 1)$ 的区间估计为

$$[\hat{\beta}_i - \text{Sd}(\hat{\beta}_i) t_{\frac{\alpha}{2}}(n-2), \quad \hat{\beta}_i + \text{Sd}(\hat{\beta}_i) t_{\frac{\alpha}{2}}(n-2)] \quad (9-2.11)$$

在R中, 由函数`lm()`可以非常方便地求出回归方程, 函数`confint()`可求出参数的置信区间. 与回归分析有关的函数还有`summary()`, `anova()`和`predict()`等. 函数`lm()`的调用格式为

`lm()`的调用格式

```
lm(formula, data, subset, weights, na.action,
```

```
method="qr", model=TRUE, x=FALSE, y=FALSE,
qr=TRUE, singular.OK = TRUE, contrasts=NULL, offset, ...)
```

说明: `formula`是显示回归模型, `data`是数据框, `subset`是样本观察的子集, `weights`是用于拟合的加权向量, `na.action`显示数据是否包含缺失值, `method`是指出用于拟合的方法, `model`, `x`, `y`, `qr`是逻辑表达, 如果是TRUE, 应返回其值. 除了第一个选项`formula`是必选项, 其它都是可选项.

函数`confint()`的调用格式为

```
confint(object, parm, level=0.95, ...)
```

说明: `object`是指回归模型, `parm`要求指出所求区间估计的参数, 默认值为所有的回归参数, `level`是指置信水平.

例 9.2.2 求例9.2.1的回归方程, 并对相应的方程作检验.

解 R程序如下:

```
> x<-c(318, 910, 200, 409, 415, 502, 314, 1210, 1022, 1225)
> y<-c(524, 1019, 638, 815, 913, 928, 605, 1516, 1219, 1624)
> lm.reg<-lm(y~1+x)
> summary(lm.reg)
> confint(lm.reg, level=0.95)
```

程序中, 第三行函数`lm()`表示使用线性回归模型 $y = \beta_0 + \beta_1 x$, 第四行函数`summary()`为提取模型计算结果. 运行结果如下:

Call:

```
lm(formula = y ~ 1 + x) # 可简化为lm(formula = y ~ x)
```

Residuals:

Min	1Q	Median	3Q	Max
-191.78	-87.05	44.75	77.86	145.66

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	395.5670	80.2611	4.929	0.00115 **
x	0.8958	0.1066	8.403	3.06e-05 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 126.6 on 8 degrees of freedom

Multiple R-Squared: 0.8982, Adjusted R-squared: 0.8855

F-statistic: 70.62 on 1 and 8 DF, p-value: 3.059e-05

结论: 从上述输出结果 p -值可以看出回归方程通过回归参数的检验与回归方程的检验, 由此得到回归方程 $Y = 395.5670 + 0.8938X$.

得到了回归方程, 还可以对误差项独立同正态分布的假设进行检验. 在R中只需再执行一个plot命令.

```
> op<-par(mfrow=c(2, 2))
> plot(lm.reg)
> par(op)
```

运行结果见图9.3. 上面的命令plot(lm.reg)实际上使用了四次plot(x, y), 产生四个图形, 它们分别为:

- 1) Residual vs fitted为拟合值 \hat{y} 对残差的图形, 可以看出, 数据点都基本均匀地分布在直线 $y = 0$ 的两侧, 无明显趋势;
- 2) Normal QQ-plot图中数据点分布趋于一条直线, 说明残差是服从正态分布的;
- 3) Scale-Location 图显示了标准化残差(standardized residuals)的平方根的分布情况. 最高点为残差最大值点;
- 4) Cook距离(Cook's distance)图显示了对回归的影响点.

■

9.2.3 预测与控制

对 $X = x_0$, $Y = y_0$ 的预测值为 $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$, 置信度为 $1 - \alpha$ 的预测区间为

$$\hat{y}_0 \pm t_{1-\alpha/2}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(\bar{X} - x_0)^2}{S_{xx}}} \quad (9-2.12)$$

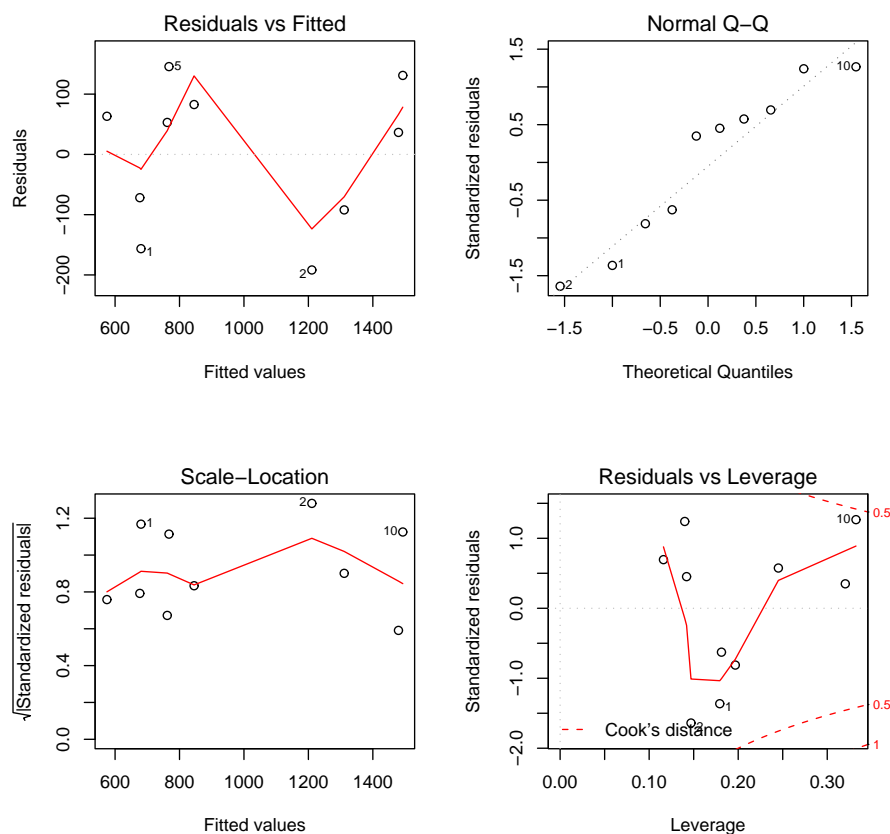


图 9.3 例9.2.2中回归分析的诊断图

由于当 $n \rightarrow \infty$ 时, $t_{1-\alpha/2}(n-2) \approx z_{1-\alpha/2}$, 于是 Y_{y_0} 的置信度为 $1-\alpha$ 的预测区间可近似为

$$[\hat{y}_0 - \hat{\sigma} z_{1-\alpha/2}, \hat{y}_0 + \hat{\sigma} z_{1-\alpha/2}]. \quad (9-2.13)$$

控制可视为是预测的反问题, 即要求观察值 Y 在某一区间 (y_l, y_u) 内取值时, 问应将 X 控制在什么范围内. 由式(9-2.13), 构造不等式

$$\begin{cases} \hat{y} - \hat{\sigma} z_{1-\alpha/2} = \hat{\beta}_0 + \hat{\beta}_1 x - \hat{\sigma} z_{1-\alpha/2} \geq y_l \\ \hat{y} + \hat{\sigma} z_{1-\alpha/2} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\sigma} z_{1-\alpha/2} \leq y_u \end{cases} \quad (9-2.14)$$

由不等式(9-2.14)得到 X 的取值范围, 并以此作为控制 X 的上下界. 为了保证得

到的控制范围有意义, y_u 和 y_l 应满足 $y_u - y_l \geq 2\hat{\sigma}z_{1-\alpha/2}$.

例 9.2.3 求例9.2.1中, $X = x_0 = 415$ 时相应 Y 的置信水平为0.95的预测区间.

解 R程序: 利用predict()函数求预测值和预测区间.

```
> point<-data.frame(x=415)
> lm.pred<-predict(lm.reg, point,
                    interval="prediction", level=0.95)
> lm.pred
      fit      lwr      upr
[1,] 765.849 457.1226 1074.575
```

程中选项interval=“prediction”表示同时要给出相应的预测区间, 选项level 指出相应的预测水平, 默认值为0.95, 这时可省略. 由计算结果得到: 当 $x = 415$ 时, y 的预测值为767.339, 预测区间为[455.5666, 1079.111]. ■

9.2.4 计算例子

例 9.2.4 表9.3是有关15个地区某种食物年需求量(X , 单位: 10吨)和地区人口增加量(Y , 单位: 千人)的资料. 利用此表数据展示一元回归模型的统计分析过程.

表 9.3 某种食物年需求量与人口增加量

编号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X	274	180	375	205	86	265	98	330	195	53	430	372	236	157	370
Y	162	120	223	131	67	169	81	192	116	55	252	234	144	103	212

计算分析过程如下:

- 1) 建立数据集, 并画出散点图: 考查数据点的分布趋势, 看是否呈直线条状分布. 程序如下

```
> x<-c(274, 180, 375, 205, 86, 265, 98, 330, 195, 53,
```

```

      430, 372, 236, 157, 370)
> y=c(162, 120, 223, 131, 67, 169, 81, 192, 116, 55,
      252, 234, 144, 103, 212)
> A<-data.frame(x, y)
> plot(A$x, A$y)

```

运行结果如图9.4所示, 可以看出, 这些点基本上(但不精确的)落在一条直线上.

2) 进行回归分析, 并在散点图上显示回归直线. R程序为

```

> lm.reg<-lm(y~x)
> summary(lm.reg)
> abline(lm.reg)

```

回归结果如下, 回归直线仍画在图9.4上.

```

Call: lm(formula = y ~ x)
Residuals:
      Min       1Q   Median       3Q      Max
-9.9610 -4.6079 -0.2618  3.1500 14.2152
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.59595     3.92745   5.753 6.67e-05 ***
x              0.53008     0.01472  36.007 2.08e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.435 on 13 degrees of freedom
Multiple R-Squared:  0.9901,
Adjusted R-squared:  0.9893
F-statistic: 1297 on 1 and 13 DF,
p-value: 2.079e-14

```

结论:

- 回归系数的估计与检验: 回归系数的估计为 $\hat{\beta}_0 = 22.59595$, $\hat{\beta}_1 = 0.53008$, 相应的标准差为 $Sd(\hat{\beta}_0) = 3.92745$, $Sd(\hat{\beta}_1) = 0.01472$. 它们的 p 值均很小, 故是非常显著的.

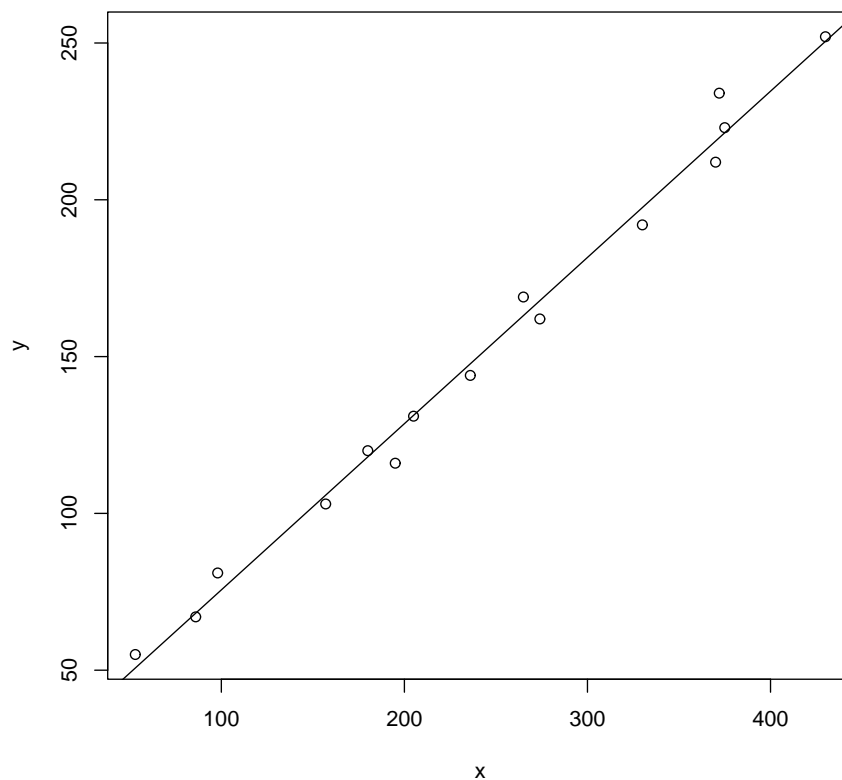


图 9.4 例9.2.4中数据的散点图

- 相关分析: 相关系数的平方 $R^2 = 0.9901$, 表明数据中99%可由回归方程来描述.
- 方程的检验: F分布的p值为 2.079×10^{-14} , 因此方程是非常显著的, 这与 R^2 的结果一致.

3) 残差分析—图形诊断: 用函数`residuals()`计算回归方程的残差, 并画出关于残差的散点图, 见图9.5.

```
> res<-residuals(lm.reg)
> plot(res)
> text(12, res[12], labels=12, adj=(.05))
```

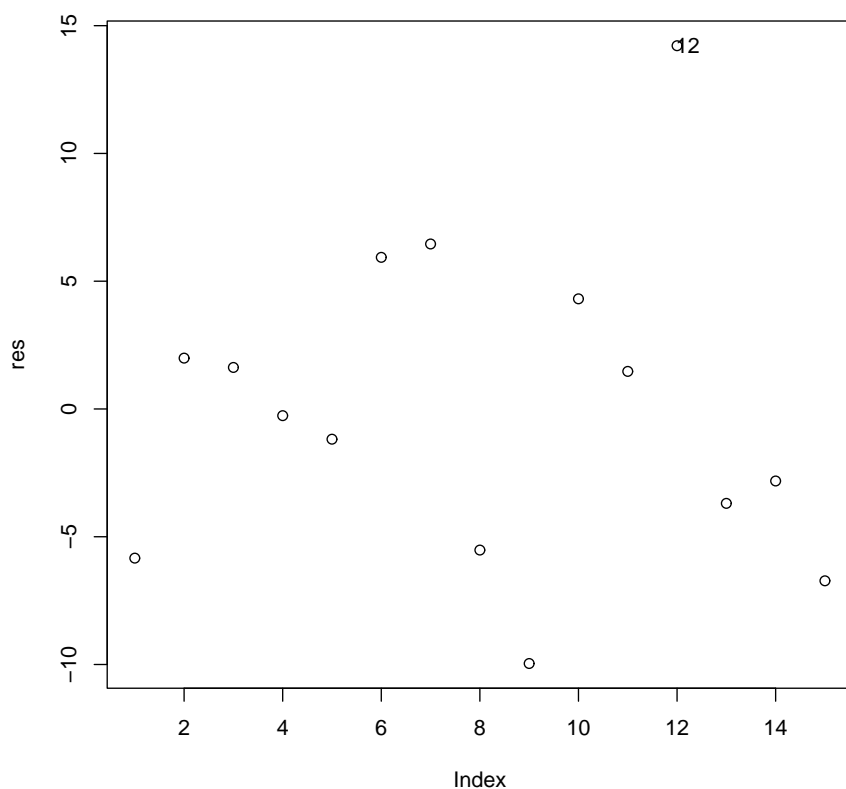



图 9.5 例9.2.4中残差的散点图

从图9.5可以看出, 第12个样本点可能有问题(程序中已用函数`text()`标注), 它比其它样本点的残差大很多, 因此, 这个点可能有问题: 或者由于模型的假设不正确, 或是 σ^2 不是常数, 或是异常点, 等等. 总之, 需要对这个问题进行进一步的分析, 这在9.4节的回归诊断中进行详细介绍.

§9.3 多元线性回归分析

许多实际问题中, 影响响应变量的因素往往不只一个而是多个, 我们称这类回归分析为多元回归分析. 这里仅讨论最为一般的线性回归问题和可以化为线性回归的问题(如本章第四节logistic回归).

9.3.1 数学模型

假设随机变量 Y 与 p 个自变量 X_1, X_2, \dots, X_p 之间存在着线性相关关系

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p,$$

其中 $\beta_0, \beta_1, \dots, \beta_p$ 是未知参数(称为回归系数或回归参数), X_1, X_2, \dots, X_p 是 p 个可以精确测量并可控制的变量(称为回归因子或预测变量), Y 为响应变量. 若其 n 次观测值为 $(X_{i1}, X_{i2}, \dots, X_{ip}, Y_i), i = 1, 2, \dots, n$, 则这 n 个观测值可写为如下形式:

$$\begin{cases} y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \dots + \beta_p X_{1p} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \dots + \beta_p X_{2p} + \varepsilon_2 \\ \vdots \\ y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \dots + \beta_p X_{np} + \varepsilon_n \end{cases} \quad (9-3.1)$$

其中 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p$ 是随机误差, 和一元线性回归分析一样, 我们假定它们相互独立且服从同一正态分布 $N(0, \sigma^2)$.

若将方程组(9-3.1)用矩阵表示, 则有

$$\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}, \quad (9-3.2)$$

其中

$$\underline{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} \dots & X_{1p} \\ 1 & X_{21} & X_{22} \dots & X_{2p} \\ \dots & \dots & \dots & \dots \\ 1 & X_{n1} & X_{n2} \dots & X_{np} \end{pmatrix}, \underline{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \underline{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

9.3.2 估计与检验

多元线性回归分析的首要任务就是通过寻求 $\underline{\beta}$ 的估计值 $\hat{\underline{\beta}}$, 建立多元线性回归方程

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_p X_p, \quad (9-3.3)$$

并对此方程及其回归系数的显著性作出检验.

与一元线性回归分析相同, 求参数 $\underline{\beta}$ 的估计值 $\hat{\underline{\beta}}$, 就是求解 β_j 使全部观察

值 Y_i 与回归值(拟合值) \hat{Y}_i ($i = 1, 2, \dots, n$)的残差平方和

$$Q(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^n (Y_i - \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip})^2 \quad (9-3.4)$$

达到最小.

可以证明, 若 \mathbf{X} 是满列秩的, 则 β 的最小二乘估计为

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \quad (9-3.5)$$

由残差向量 $\hat{\epsilon} = \mathbf{Y} - \mathbf{X}\hat{\beta}$, 通常取

$$\hat{\sigma}^2 = \frac{\hat{\epsilon}'\hat{\epsilon}}{n-p-1} \quad (9-3.6)$$

作为 σ^2 的估计, 也称为 σ^2 的最小二乘估计.

得到了回归方程后, 由于我们无法像一元线性回归分析那样用直观的方法帮助判断 Y 与 X_1, X_2, \dots, X_p 之间是否有线性关系, 为此必须对回归方程进行显著性检验. 其次在 p 个变量中, 每个自变量对 y 的影响程度是不同的, 甚至有的自变量是可有可无的. 这表现在回归系数中有的绝对值很大, 有的很小或接近于零, 这就需要对回归系数进行显著性检验.

回归方程显著性检验

考虑假设检验问题:

$$H_0: \beta_0 = \beta_1 = \dots = \beta_p = 0 \leftrightarrow H_1: \beta_0, \beta_1, \dots, \beta_p \text{不全为} 0$$

可以证明当 H_0 成立时, 统计量

$$F = \frac{SS_R/p}{SS_E/(n-p-1)} \sim F(p, n-p-1) \quad (9-3.7)$$

其中

$$SS_R = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2, \quad SS_E = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2,$$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \dots + \hat{\beta}_p X_{ip}.$$

称 SS_R 为回归平方和, 称 SS_E 为残差平方和.

因此, 对于给定的显著性水平 α , 检验回归方程的拒绝域为

$$F > F_{1-\alpha}(p, n-p-1).$$

回归系数的显著性检验

$$H_{0j} : \beta_j = 0 \leftrightarrow H_{1j} : \beta_j \neq 0 (j = 0, 1, \dots, p)$$

可以证明, 当 H_{0j} 成立时, 有

$$F_j = \frac{Q_j}{SS_E/(n-p-1)} \sim F(1, n-p-1) \quad (9-3.8)$$

其中 $Q_j = SS_{E(j)} - SS_E$, $SS_{E(j)}$ 是去掉 X_j 后的残差平方和.

对于给定的显著性水平 α , 当 $F_j > F_{1-\alpha}(1, n-p-1)$ 时拒绝 H_{0j} , 认为变量 X_j 对 Y 有显著影响.

例 9.3.1 某公司在各地区销售一种特殊的化妆品, 该公司观测了15个城市在某月对该化妆品的销售量(Y), 使用该化妆品的人数(X_1)和人均收入(X_2), 数据见表9.4. 试建立 Y 与 X_1, X_2 的线性回归方程, 并作相应的检验.

解 R程序为:

```
> y<-c(162, 120, 223, 131, 67, 169, 81, 192, 116, 55,
      252, 232, 144, 103, 212)
> x1<-c(274, 180, 375, 205, 86, 265, 98, 330, 195, 53,
      430, 372, 236, 157, 370)
> x2<-c(2450, 3250, 3802, 2838, 2347, 3782, 3008, 2450,
      2137, 2560, 4020, 4427, 2660, 2088, 2605)
> sales<-data.frame(y, x1, x2)
> lm.reg<-lm(y~x1+x2, data=sales)
> summary(lm.reg)
```

运行结果:

Call:

表 9.4 某种化妆品的销售量及有关指标

地区 i	销售量(Y)/箱	人数(X_1)/千人	人均收入(X_2)/元
1	162	274	2450
2	120	180	3250
3	223	375	3802
4	131	205	2838
5	67	86	2347
6	167	265	3782
7	81	98	3008
8	192	330	2450
9	116	195	2137
10	55	53	2560
11	252	430	4020
12	232	372	4427
13	144	236	2660
14	103	157	2088
15	212	370	2605

```
lm(formula = y ~ x1 + x2, data = sales)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.8312	-1.2063	-0.2436	1.4819	3.3025

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.4457284	2.4266934	1.420	0.181
x1	0.4959724	0.0060455	82.039	< 2e-16 ***
x2	0.0092049	0.0009668	9.521	6.07e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.173 on 12 degrees of freedom

Multiple R-squared: 0.9989, Adjusted R-squared: 0.9988

F-statistic: 5699 on 2 and 12 DF, p-value: < 2.2e-16

结论: 由于用于回归方程检验的 F 统计量的 p 值与用于回归系数检验的 t 统计量的 p 值均很小(<0.05), 因此回归方程与回归系数的检验都是显著的. 回归方程为

$$Y = 3.4457 + 0.4960X_1 + 0.0092X_2.$$

■

9.3.3 预测与控制

当多元线性回归方程经过检验是显著的, 且其中每一个回归系数均显著时(不显著的先剔除), 这时可用此回归方程作预测.

给定 $\underline{X} = \underline{x}_0 = (x_{01}, x_{02}, \dots, x_{0p})^T$, 将其代入回归方程, 得预测值

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \dots + \hat{\beta}_p x_{0p}. \quad (9-3.9)$$

相应的置信度为 $1 - \alpha$ 的预测区间为

$$\hat{Y}_0 \pm t_{1-\alpha/2}(n-p-1)\hat{\sigma}\sqrt{1 + \underline{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\underline{x}_0}. \quad (9-3.10)$$

例 9.3.2 求例9.3.1中 $\underline{X} = \underline{x}_0 = (200, 3000)'$ 时相应 Y 的观测值与0.95预测区间.

解 与一元回归一样, 在 \mathbf{R} 中仍使用函数predict()求多元回归预测.

```
> exa<-data.frame(x1=200, x2=3000)
> lm.pred<-predict(lm.reg, exa, interval="prediction", level=0.95)
> lm.pred
```

运行结果:

```
fit      lwr      upr
[1,] 130.2549 125.3274 135.1824
```

由此求得, $\hat{Y}_0 = 130.2549$, 相应的 Y 的0.95的预测区间为[125.3274, 135.1824].

■

9.3.4 计算例子

例 9.3.3 27名糖尿病人的血清总胆固醇(X_1)、甘油三酯(X_2)、空腹胰岛素(X_3)、糖化血红蛋白(X_4)、空腹血糖(Y)的测量值列于表9.5中, 试建立血糖与其它指标的多元线性回归方程, 并作进一步分析.

解 计算分析过程及相应的R程序如下:

1) 建立数据集:

```
> y<-c(11.2, 8.8, 12.3, 11.6, 13.4, 18.3, 11.1, 12.1,
        9.6, 8.4, 9.3, 10.6, 8.4, 9.6, 10.9, 10.1,
        14.8, 9.1, 10.8, 10.2, 13.6, 14.9, 16.0, 13.2,
        20.0, 13.3, 10.4)
> x1<-c(5.68, 3.79, 6.02, 4.85, 4.60, 6.05, 4.90, 7.08,
        3.85, 4.65, 4.59, 4.29, 7.97, 6.19, 6.13, 5.71,
        6.40, 6.06, 5.09, 6.13, 5.78, 5.43, 6.50, 7.98,
        11.54, 5.84, 3.84)
> x2<-c(1.90, 1.64, 3.56, 1.07, 2.32, 0.64, 8.50, 3.00,
        2.11, 0.63, 1.97, 1.97, 1.93, 1.18, 2.06, 1.78,
        2.40, 3.67, 1.03, 1.71, 3.36, 1.13, 6.21, 7.92,
        10.89, 0.92, 1.20)
> x3<-c(4.53, 7.32, 6.95, 5.88, 4.05, 1.42, 12.60, 6.75,
        16.28, 6.59, 3.61, 6.61, 7.57, 1.42, 10.35, 8.53,
        4.53, 12.79, 2.53, 5.28, 2.96, 4.31, 3.47, 3.37,
        1.20, 8.61, 6.45)
> x4<-c(8.2, 6.9, 10.8, 8.3, 7.5, 13.6, 8.5, 11.5,
        7.9, 7.1, 8.7, 7.8, 9.9, 6.9, 10.5, 8.0,
        10.3, 7.1, 8.9, 9.9, 8.0, 11.3, 12.3, 9.8,
        10.5, 6.4, 9.6)
> blood<-data.frame(y, x1, x2, x3, x4)
```

2) 建立多元线性回归方程:

表 9.5 27名糖尿病人的指标

i	X_1	X_2	X_3	X_4	Y
1	5.68	1.90	4.53	8.2	11.2
2	3.79	1.64	7.32	6.9	8.8
3	6.02	3.56	6.95	10.8	12.3
4	4.85	1.07	5.88	8.3	11.6
5	4.60	2.32	4.05	7.5	13.4
6	6.05	0.64	1.42	13.6	18.3
7	4.90	8.50	12.60	8.5	11.1
8	7.08	3.00	6.75	11.5	12.1
9	3.85	2.11	16.28	7.9	9.6
10	4.65	0.63	6.59	7.1	8.4
11	4.59	1.97	3.61	8.7	9.3
12	4.29	1.97	6.61	7.8	10.6
13	7.97	1.93	7.57	9.9	8.4
14	6.19	1.18	1.42	6.9	9.6
15	6.13	2.06	10.35	10.5	10.9
16	5.71	1.78	8.53	8.0	10.1
17	6.40	2.40	4.53	10.3	14.8
18	6.06	3.67	12.79	7.1	9.1
19	5.09	1.03	2.53	8.9	10.8
20	6.13	1.71	5.28	9.9	10.2
21	5.78	3.36	2.96	8.0	13.6
22	5.43	1.13	4.31	11.3	14.9
23	6.50	6.21	3.47	12.3	16.0
24	7.98	7.92	3.37	9.8	13.2
25	11.54	10.89	1.20	10.5	20.0
26	5.84	0.92	8.61	6.4	13.3
27	3.84	1.20	6.45	9.6	10.4

```
> lm.reg<-lm(y~x1+x2+x3+x4, data=blood)
> summary(lm.reg)
```


运行结果如下:

```
Call: lm(formula = y ~ x1 + x2 + x3 + x4, data = blood)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-3.6268	-1.2004	-0.2276	1.5389	4.4467

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	5.9433	2.8286	2.101	0.0473 *
x1	0.1424	0.3657	0.390	0.7006
x2	0.3515	0.2042	1.721	0.0993 .
x3	-0.2706	0.1214	-2.229	0.0363 *
x4	0.6382	0.2433	2.623	0.0155 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.01 on 22 degrees of freedom

Multiple R-Squared: 0.6008,

Adjusted R-squared: 0.5282

F-statistic: 8.278 on 4 and 22 DF,

p-value: 0.0003121

结论: 回归方程的系数的显著性不高, 有的甚至没有通过检验(X_1 与 X_2), 这说明如果选择全部变量构造方程, 效果并不好. 这就涉及到变量选择的问题, 以建立“最优”的回归方程.

- 3) 变量选择与最优回归: **R**软件提供了获得“最优”回归方程的方法, “逐步回归法”的计算函数`step()`, 它是以Akaike信息统计量为准则(简称AIC准则), 通过选择最小的AIC信息统计量, 来达到删除或增加变量的目的. 函数`step()`的调用格式为

`step()`函数的调用格式

```
step(object, scope, scale=0,
      direction=c("both", "backward", "forward",
                  trace=1, keep=NULL, steps=1000, k=2, ...)
```

说明: `object`是线性模型或广义线性模型分析的结果, `scope`是确定逐步搜索的区域, `direction`确定逐步搜索的方向: “both”是“一切子集回归

法”，“backward”是“向后法”，“forward”是向前法，默认值为both. 其它参数见在线帮助.

对于本例用函数step()作逐步回归:

```
> lm.step<-step(lm.reg)
```

回归结果为:

```
Start:  AIC= 42.16
y ~ x1 + x2 + x3 + x4
      Df Sum of Sq    RSS    AIC
- x1    1     0.613  89.454  40.343
<none>                88.841  42.157
- x2    1    11.963 100.804  43.568
- x3    1    20.064 108.905  45.655
- x4    1    27.794 116.635  47.507
```

```
Step:  AIC= 40.34
y ~ x2 + x3 + x4
      Df Sum of Sq    RSS    AIC
<none>                89.454  40.343
- x3    1    25.690 115.144  45.159
- x2    1    26.530 115.984  45.356
- x4    1    32.269 121.723  46.660
```

结论: 用全部变量作回归方程时, AIC统计量的值为42.16, 如果去掉变量 X_1 , AIC统计量的值为40.34; 如果去掉变量 X_2 , AIC统计量的值为43.568, 依次类推. 由于去掉 X_1 使AIC统计量达到最小, 因此R软件会自动去掉变量 X_1 , 进入下一轮计算. 在下一轮中, 无论去掉哪一个变量, AIC统计量的值均会升高, 因此R软件自动终止计算, 得到“最优”回归方程.

再用函数summary()提取相关回归信息.

```
> summary(lm.step)
```

提取结果为:

```

Call:
lm(formula = y ~ x2 + x3 + x4, data = blood)

Residuals:
    Min       1Q   Median       3Q      Max
-3.2692 -1.2305 -0.2023  1.4886  4.6570

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.4996     2.3962   2.713  0.01242 *
x2             0.4023     0.1541   2.612  0.01559 *
x3            -0.2870     0.1117  -2.570  0.01712 *
x4             0.6632     0.2303   2.880  0.00845 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.972 on 23 degrees of freedom
Multiple R-squared:  0.5981,    Adjusted R-squared:  0.5456
F-statistic: 11.41 on 3 and 23 DF,  p-value: 8.793e-05

```

结论: 回归系数的显著性水平有很大提高, 所有的检验均是显著的, 由此得到“最优”的回归方程

$$Y = 6.4996 + 0.4023X_2 - 0.2870X_3 + 0.6632X_4.$$



§9.4 回归诊断

前面介绍了如何得到回归模型, 但没有对回归模型的一些特性作进一步的研究, 并且没有研究对回归模型产生较大影响的异常值问题. 异常值的存在往往会给回归模型带来不稳定, 为此, 人们提出了所谓回归诊断的问题, 其主要内容有: 残差分析、影响分析、共线性诊断等.

9.4.1 残差分析

残差及残差图

残差向量 $\hat{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$ 是模型中误差项 ε 的估计, 其中 $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$ 称为帽子矩阵. 由于

$$E(\hat{\varepsilon}) = 0, \quad \text{Var}(\hat{\varepsilon}) = \sigma^2(\mathbf{I} - \mathbf{H}) \quad (9-4.1)$$

因此, 对每个 $\hat{\varepsilon}_i$, 有

$$\frac{\hat{\varepsilon}_i}{\sigma\sqrt{1-h_{ii}}} \sim N(0, 1), \quad (9-4.2)$$

其中 h_{ii} 是矩阵 \mathbf{H} 对角线上第 i 个元素.

当用 $\hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-p-1}$ 去估计 σ^2 时, 称

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \quad (9-4.3)$$

为标准化残差, 或内学生化残差.

当用 $\hat{\sigma}_{(i)}^2 = \frac{1}{n-p-2} \sum_{j \neq i} (Y_j - \tilde{X}_j' \hat{\beta}_{(i)})^2$ 去估计 σ^2 时, 称

$$\frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} \quad (9-4.4)$$

为学生化残差, 或外学生化残差. 其中 $\hat{\beta}_{(i)}$ 是删去第 i 个样本点后用余下的 $n-1$ 个样本点求得的回归系数, \tilde{X}_j 为设计矩阵 \mathbf{X} 的第 j 行.

R 软件中, 分别用函数 `residuals()`, `rstandard()` 和 `rstudent()` 来计算残差、标准化残差和学生化残差. 这些函数的调格式分别为:

`residuals()` 的调用格式

```
residuals(object, ...)
resid(object, ...)
```

`rstandard()` 的调用格式

```
rstandard(model, infl=lm.influence(model, do.coef=FALSE),
```

```
sd=sqrt(deviance(model)/df.residual(model)), ...)
```

rstudent()的调用格式

```
rstudent(model, infl=lm.influence(model, do.coef=FALSE,
res=infl$wt.res, ...)
```

说明: object或model是由线性模型函数lm()或广义线性模型函数glm()生成的对象, infl是由lm.influence()返回值得到的影响结构, sd是模型的标准差, res是模型残差.

凡是以残差为纵坐标, 以观测值 Y_i , 预测值 \hat{Y}_i , 自变量 $X_{ij}(j = 1, 2, \dots, m)$ 或序号、观测时间等为横坐标的散点图, 均称为残差图. 如果多元线性回归模型的假定成立, 从理论上可证明 r_1, r_2, \dots, r_n 相互独立且近似服从 $N(0, 1)$, 故关于观测值等的残差图中散点应随机的分布在-2到+2的带子里, 并称之为正常残差图(见图9.6a), 否则称为异常残差图(见图9.6b,c,d).

例 9.4.1 计算例9.3.1的残差和标准残差, 并画出相应的残差散点图.

解 R程序为:

```
> y.res<-residuals(lm.reg) #计算残差
> print(y.res)
> y.rst<-rstandard(lm.reg) #计算标准化残差
> print(y.rst)
> y.fit<-predict(lm.reg) #计算预测值
> op<-par(mfrow=c(1, 2)) #将两张散残差点图一并输出
> plot(y.res~y.fit)
> plot(y.rst~y.fit)
> par(op)
```

计算结果如下, 图形见图9.7. 从图9.7可以看出, 残差具有相同的分布且满足模型的各个假设条件.

残差:

1	2	3	4	5
0.1058453	-2.6366596	-1.4323818	-0.2435552	-0.7032355
6	7	8	9	10

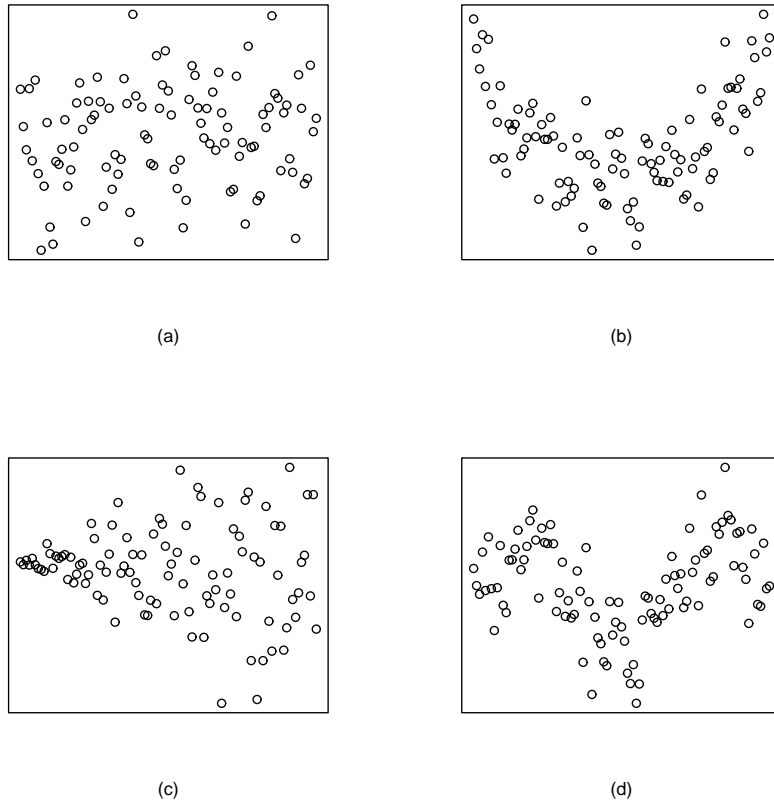


图 9.6 (1)正常的残差图; (b)应改为曲线模型; (c)主差齐性不成立; (d)观测值不独立

```

-0.6913175  1.2606626  2.3313896 -3.8312025  1.7032127
          11          12          13          14          15
-1.7175312  3.3024787 -0.9802297  2.4667892  1.0657346
标准化残差:
          1          2          3          4          5
0.05281317 -1.30635637 -0.73052549 -0.11643248 -0.36046378
          6          7          8          9         10
-0.35064339 0.66372152  1.23228395 -1.92770717  0.91558703
          11          12          13          14          15
-0.93261640 1.89069180 -0.47083133  1.24503836  0.57927692

```

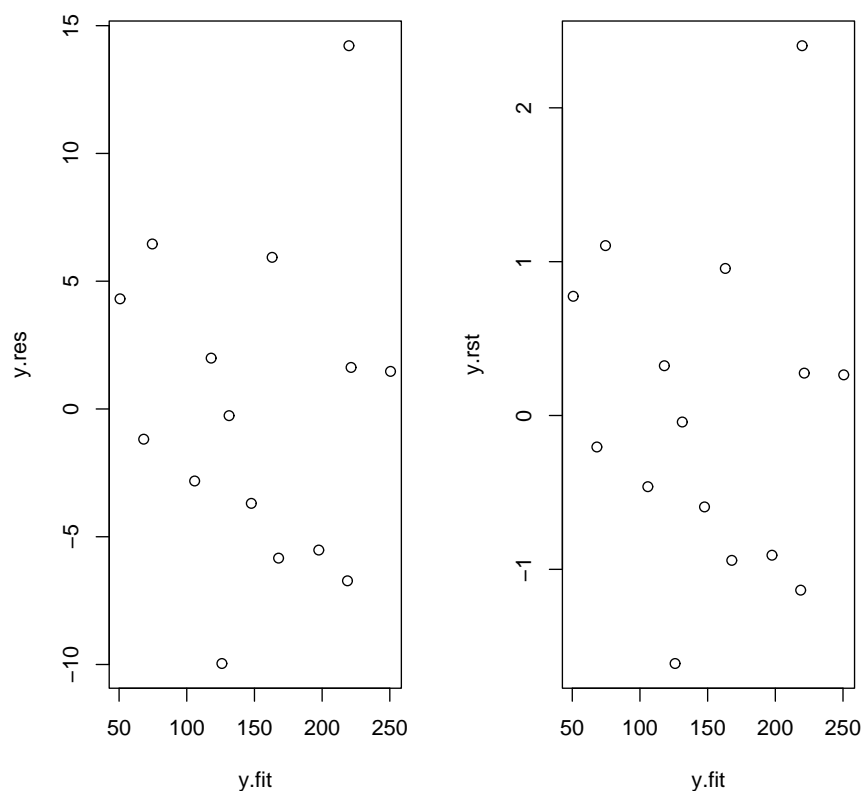


图 9.7 例9.4.1中的残差与标准化残差图



方差齐性的诊断及修正方法

从图9.7的残差图可以看出, 当残差的绝对值随预测值的增加也有明显增加的趋势(或减少的趋势, 或先增加后减少的趋势)时, 表示关于误差的方差齐性(即误差方差 $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_n^2$)的假定不成立.

误差方差非齐性时, 有时可以通过对因变量作适当的变换, 即令 $Z = f(Y)$, 使得关于因变量 Z 在回归中误差的方差接近齐性. 理论上根据观测向量 Y 的性质(如均值 $E(Y)$ 和方差 $\text{Var}(Y)$ 的关系等)可以判断出应做什么样的变换合适. 实用上, 常选用一些常用的变换, 变换后重新做回归及残差图, 如残差图有改

善或已属正常, 则认为该变换是合适的; 否则改变变换函数重新计算, 直到找到合适的变换, 常见的方差稳定性变换有:

- 1) 开方变换 $Z = \sqrt{Y}$ ($Y > 0$);
- 2) 对数变换 $Z = \ln(Y)$ ($Y > 0$);
- 3) 倒数(逆)变换 $Z = 1/Y$ ($Y \neq 0$);
- 4) BOX-COX变换 $Z = \frac{X^\lambda - 1}{\lambda}$.

$\lambda = 0$ 时的BOX-COX变换即为对数变换.

例 9.4.2 在对27家企业单位的研究中, 记录了企业管理人数(Y)与工人人数(X)资料(见表9.6). 试建立 Y 对 X 的回归方程.

表 9.6 27各企业单位中企业管理人员数与员工数

序号	X	Y	序号	X	Y	序号	X	Y
1	294	50	10	697	78	19	700	106
2	247	40	11	688	80	20	850	128
3	267	45	12	630	84	21	980	130
4	358	55	13	709	88	22	1025	160
5	423	70	14	627	97	23	1021	97
6	311	65	15	615	100	24	1200	180
7	450	55	16	999	109	25	1250	112
8	534	62	17	1022	114	26	1500	210
9	438	68	18	1015	117	27	1650	135

解 分析过程如下:

- 1) 输入数据.

```
> x<-c(294, 247, 267, 358, 423, 311, 450, 534, 438, 697,
      688, 630, 709, 627, 615, 999, 1022, 1015, 700, 850,
```



```

      980, 1025, 1021, 1200, 1250, 1500, 1650)
> y<-c(50, 40, 45, 55, 70, 65, 55, 62, 68, 78,
      80, 84, 88, 97, 100, 109, 114, 117, 106, 128,
      130, 160, 97, 180, 112, 210, 135)
> persons<-data.frame(x, y)

```

2) 作线性回归模型.

```

> lm.reg<-lm(y~x)
> summary(lm.reg)

```

得到:

Call:

```
lm(formula = y ~ x)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-47.645	-11.136	-4.278	11.683	41.677

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	25.09434	9.27542	2.705	0.0121 *
x	0.09549	0.01099	8.691	5.02e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.08 on 25 degrees of freedom

Multiple R-squared: 0.7513, Adjusted R-squared: 0.7414

F-statistic: 75.54 on 1 and 25 DF, p-value: 5.018e-09

显然, 回归系数和回归方程都通过了检验, 所以Y对X的一元回归方程为

$$Y = 25.09434 + 0.0549X.$$

3) 回归诊断. 画出标准化残差散点图, **R**程序为

```

> y.rst<-rstandard(lm.reg)
> y.fit<-predict(lm.reg)
> plot(y.rst~y.fit)

```

其图形如图9.8图所示. 直观上容易看出, 残差图从左向右逐渐散开, 这是方差齐性不成立的典型征兆. 所以, 应考虑对响应变量 Y 作变换.

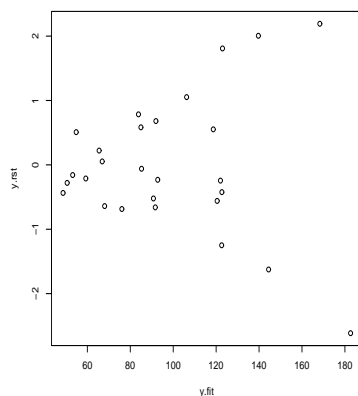
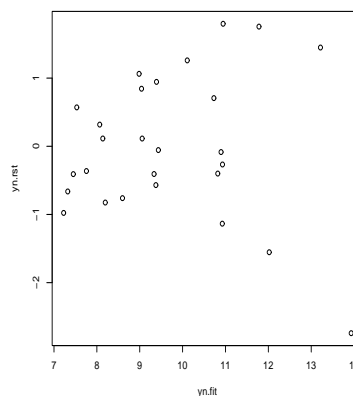


图 9.8 标准化残差图

图 9.9 对 Y 开方运算后标准化残差图

- 4) 模型更新. 在新的平方变换下进行回归分析, 并进行回归诊断, 相应的R程序为:

```

> lm.new_reg<-update(lm.reg, sqrt(.)~.)
> coef(lm.new_reg)

```

说明: 函数`update()`对回归模型按给定的方差稳定化变换进行修正, 函数`coef()`提取回归系数的估计.

计算结果为:

```

(Intercept)          x
6.044644223 0.004780664

```

由此得到新的回归方程为

$$Y = (6.044644223 + 0.004780664X)^2,$$

即

$$Y = 36.5377238 + 0.0577948X + 2.28547 \times 10^{-5}X^2.$$

最后画出变换后的标准化残差散点图, 程序为

```
> yn.rst<-rstandard(lm.new_reg)
> yn.fit<-predict(lm.new_reg)
> plot(yn.rst~yn.fit)
```

其图形如图9.9所示, 散点图的趋势大有改善.

■

异常点的识别

如果拟合后的模型能够很好地描述这组数据, 那么残差对预测值的散点图应该像一些随机散布的点. 可是, 若某个观测不能和其它数据一起用这个模型表示, 那么那个观测的残差通常很大. 这里“很大”指的是残差的绝对值. 因为一个“很大”的残差可能是正的也可能是负的. 如果只有占很小百分比的观测出现大的残差, 那么这些观测可能是异常点(outliers) — 它们不能用来与其余数据一起拟合模型. 因此对数据中有残差“很大”的观测点, 必须仔细地检查.

一般把标准化残差的绝对值 ≥ 2 的观测点认为是可疑点; 而标准化残差的绝对值 ≥ 3 的观测点认为是异常点.

例 9.4.3 对例9.3.1中得到回归方程, 判断是否有异常点.

解 由例的计算结果并结合图形可以看出, 第12个点的残差比较大, 被认定为异常点. 它可以用下列语句将异常点标出(见图9.10).

```
> text(219.78476, 2.4037012, labels=12, adj=(.2))
```

这里再做一个简单处理, 去掉第12观测样本点, 并重复上述回归分析及残差分析的过程, 得到新的标准化残差图9.11. 与图9.10相比, 现在残差点的分布已有了很大的改进, 它们基本上落在 $[-2, 2]$ 的带状区域内. 但好像仍有一个可疑点存在, 故需进一步分析(从略).

■

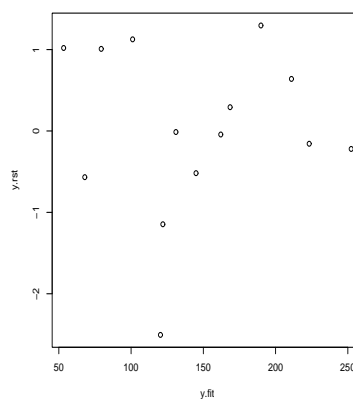
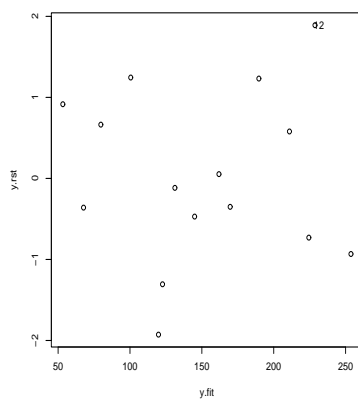


图 9.10 标准化残差图: 全部数据 图 9.11 标准化残差图: 去掉12号观测

9.4.2 影响分析

从分析观测点对回归结果的影响入手, 找出对回归结果影响很大的观测点的分析方法称为影响分析.

影响函数

称向量 $\mathbf{F}_i = \hat{\beta}_{(i)} - \hat{\beta}$ 为第 i 个观测点的影响函数 ($i = 1, 2, \dots, n$), 其中 $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_m)'$ 是回归模型中参数向量 β 的最小二乘估计; $\hat{\beta}_{(i)}$ 是去掉第 i 个观测点重新计算得出 β 的最小二乘估计. 直观地看, 若 $\hat{\beta}$ 与 $\hat{\beta}_{(i)}$ 相差较大, 则表明第 i 个观测点对回归结果的影响就大.

R 软件中计算影响函数的函数为 `lm.influence()`, 其调用格式为

—— `lm.influence()` 的调用格式 ——

```
lm.influence(model, do.coef=TRUE)
```

说明: `model` 为回归模型. `do.coef=TRUE` 表示结果要求给出去掉第 i 个观测点后的模型回归系数.

Cook距离

Cook距离是从估计角度提出的一种度量第*i*个观测点对回归影响大小的统计量. 对每一个观测点, 定义Cook距离为

$$D_i(M, C_0) = \frac{(\hat{\beta}_{(i)} - \hat{\beta})' M (\hat{\beta}_{(i)} - \hat{\beta})}{C_0} \quad (9-4.5)$$

一般取*M*为观测数据的离差阵, *C*₀为回归模型均方误差(EMS). *D_i*大的观测点称为强影响点. 一般建议使用的判别标准是: 当 $|D_i| > 4/n$ 时, 认为是强影响点, 其中*n*是样本容量.

R软件中用于计算Cook距离的函数为`cooks.distance()`, 其调用格式为:

—— `cooks.distance()` 的调用格式 ——

```
cooks.distance(model, infl=im.influence(model, do.coef=FALSE),
               res=weighted.residuals(model),
               sd=sqrt(deviance(model)/df.residual(model)),
               hat=infl$hat, ...)
```

DFFITS准则

Belsley, Kuh和Welsch(1980)给出另一种准则, 所用的统计量为

$$D_i(\sigma) = \sqrt{\frac{h_{ii}}{1 - h_{ii}}} \frac{\hat{\varepsilon}_i}{\sigma \sqrt{1 - h_{ii}}} \quad (9-4.6)$$

其中 σ 用估计量 $\hat{\sigma}(i)$ 来代替. 对于第*i*个样本, 如果有

$$|D_i(\sigma)| > 2\sqrt{\frac{p+1}{n}},$$

则认为第*i*个样本的影响比较, 这里的*p* + 1是参数向量 β 的维数, *n*是样本容量.

R软件给出了DFFITS准则的计算函数`dffits()`, 其调用格式为:

—— `dffits()` 的调用格式 ——

```
dffits(model, infl=..., res=...)
```

COVRATIO准则

利用全部样本点的回归系数估计值的协方差阵和去掉第*i*个样本点后回归系数估计值的协方差阵分别为

$$\text{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}, \quad \text{Var}(\hat{\beta}_{(i)}) = \sigma^2(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}$$

其中 $\mathbf{X}_{(i)}$ 是 \mathbf{X} 剔除第*i*行后得到的矩阵. 使用时分别用 $\hat{\sigma}$ 和 $\hat{\sigma}_{(i)}$ 替代上面两式中的 σ .

为了比较其回归系数的精度, 考虑这两个协方差阵行列式的比值

$$\begin{aligned} \text{COVRATIO}_i &= \frac{\det(\hat{\sigma}_{(i)}^2(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1})}{\det(\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1})} \\ &= \frac{(\hat{\sigma}_{(i)}^2)^{p+1}}{(\hat{\sigma}^2)^{p+1}} \cdot \frac{1}{1 - h_{ii}}, \quad i = 1, 2, \dots, n \end{aligned} \quad (9-4.7)$$

如果有一个样本点所对应的COVRATIO值离1越远, 则认为该样本点影响越大.

R软件中计算COVRATIO值的函数为`covratio()`, 其调用格式为:

covration() 的调用格式

```
covratio(model, infl=lm.influence(model, do.coef=FALSE),
          res=weighted.residuals(model))
```

注: 上面介绍了四种分析强影响点的方法及相应的**R**函数, 每种方法找到的点是否是真正的强影响点还需要根据具体情况进行分析. 在**R**软件中, 函数`influence.measures()`可以做回归诊断中影响分析的概括, 它的调用格式为:

influence.measures() 的调用格式

```
influence.measures(model)
```

结果返回一个列表, 列表中包括DFFITS统计量, COVRATIO统计量, Cooks距离等.

例 9.4.4 电影院老板调查电视广告的费用 x_1 和报纸广告的费用 x_2 对每周总收入 y 的影响(单位: 元), 数据见表9.7. 试给出回归分析, 并进行回归诊断.

表 9.7 电视广告和报纸广告费用与收入的数据

x_1	x_2	y	x_1	x_2	y
1500	5000	96000	2000	2000	90000
1500	4000	95000	2500	2500	92000
3300	3000	95000	2300	3500	95000
4200	2500	94000	2500	3000	94000

解 R程序如下:

```
> x1<-c(1500, 1500, 3300, 4200, 2000, 2500, 2300, 2500)
> x2<-c(5000, 4000, 3000, 2500, 2000, 2500, 3500, 3000)
> y<-c(96000, 95000, 95000, 94000, 90000, 92000, 95000, 94000)
> money<-data.frame(x1, x2, y)
> lm.reg<-lm(y~x1+x2, data=money)
> summary(lm.reg)
> influence.measures(lm.reg)
```

这里只给出influence.measures()函数语句的返回结果:

Influence measures of

lm(formula = y ~ x1 + x2, data = money) :

```
      dfb.1_   dfb.x1  dfb.x2  dffit  cov.r cook.d   hat inf
1  1.98495  0.06107 -3.8115 -5.167 0.0461 2.2838 0.633  *
2  0.11517 -0.26496  0.1105  0.533 1.7687 0.1018 0.301
3 -0.19968  0.33167  0.1285  0.468 1.7188 0.0791 0.262
4  0.88978 -1.56524 -0.3538 -1.871 1.8860 1.0056 0.660  *
5 -1.53634  1.11782  1.4109 -1.633 2.1558 0.8133 0.645  *
6 -0.16907  0.08580  0.1621 -0.242 2.1804 0.0233 0.226
7 -0.00772 -0.00518  0.1016  0.383 1.2356 0.0499 0.140
8  0.10067 -0.03095 -0.0686  0.305 1.4686 0.0335 0.132
```

结论: 可以看出, 第1、4、5个观测点为强影响点, 结果中已用“*”号标出. 其中, 第一个样本点的cook.d值为2.2838比 $4/n=4/8=0.5$ 大得多; 第四个样本点

的cov.r值为1.8860, 与1距离很远; 第五个样本点的dffit值的绝对值1.633明显大于 $2\sqrt{\frac{p+1}{n}} = 2\sqrt{\frac{2}{8}} = 1$. 故这三个点被认为是强影响点. ■

9.4.3 共线性诊断

共线性问题是指拟合多元线性回归时, 自变量之间存在线性关系或近似线性关系. 自变量之间的线性关系将会隐藏变量的显著性, 增加参数估计的误差, 还会产生一个很不稳定的模型. 所以, 共线性诊断就是找出哪些变量间存在共线关系, 主要有以下几种方法:

特征值法

首先把 $\mathbf{X}'\mathbf{X}$ 变换为主对角线是1的矩阵, 然后求特征值和特征向量. 若有 r 个特征值近似等于0, 则回归设计阵 \mathbf{X} 中有 r 个共线性关系, 且共线性关系的系数向量就是近似为0的特征值对应的特征向量.

R软件中提供了计算矩阵特征值和特征向量的函数为`eigen()`, 其调用格式为:

`eigen()`的调用格式

```
eigen(x, symmetric, only.values=FALSE, EISPACK=FALSE)
```

说明: x 为所求矩阵, `symmetric`规定矩阵的对称性, `only.value=TRUE`表示只返回了特征值. 否则, 返回特征值和特征向量. 其它参数见在线帮助.

条件指数

若自变量的交叉乘积矩阵 $\mathbf{X}'\mathbf{X}$ 的特征值为 $d_1^2 \geq d_2^2 \geq \dots \geq d_k^2$, 则 \mathbf{X} 的条件数 d_1/d_k 就是刻画矩阵的奇异性的一个指标, 故称 $d_1/d_j (j = 1, \dots, k)$ 为条件指数.

一般认为, 若条件指数值在10与30之间为弱相关; 在30与100之间为中等相关; 大于100表明有强相关性.

在**R**软件中, 可使用函数`kappa()`计算矩阵的条件数, 其调用格式为:

`kappa()`的调用格式

```
kappa(x, exact=FALSE, ...)
```


说明: x 是矩阵, **exact**是逻辑变量: 当**exact**=TRUE时, 精确计算条件数; 否则近似计算条件数.

方差膨胀因子

方差膨胀因子VIF是指回归系数的估计量由于自变量共线性使得方差增加的一个相对度量. 对第 j 个回归系数($j = 1, 2, \dots, m$), 它的方差膨胀因子定义为

$$\begin{aligned} \text{VIF}_j &= \frac{\text{第}j\text{个回归系数的方差}}{\text{自变量不相关时第}j\text{个回归系数的方差}} \\ &= \frac{1}{1 - R_j^2} = \frac{1}{\text{TOL}_j} \end{aligned} \quad (9-4.8)$$

其中 $1 - R_j^2$ 是自变量 x_j 对模型中其余自变量线性回归模型的 R 平方, VIF_j 的倒数 TOL_j 也称容限(Tolerance).

一般建议: 若 $\text{VIF} > 10$, 表明模型中有很强的共线性问题.

R软件的**DAAG**程序包中, 函数**vif()**用来计算方差膨胀因子, 其调用格式为:

vif()的调用格式

```
vif(lmobj, digits=5)
```

说明: **lmobj**为由**lm()**生成的对象, **digits**给出小数点位数, 缺省为5位.

例 9.4.5 某种水泥在凝固时单位质量所释放的热量为 Y 卡/克, 它与水泥中下列四种化学成分有关:

X_1 — $3\text{CaO} \cdot \text{Al}_2\text{O}_3$ 的成分(%)

X_2 — $3\text{CaO} \cdot \text{SiO}_2$ 的成分(%)

X_3 — $4\text{CaO} \cdot \text{Al}_2\text{O}_3 \cdot \text{Fe}_2\text{O}_3$ 的成分(%)

X_4 — $2\text{CaO} \cdot \text{SiO}_2$ 的成分(%)

共观测了13组数据(见表9.8), 试对自变量的共线性进行诊断.

解 回归分析的**R**程序如下:

```
> x1<-c(7, 1, 11, 11, 7, 11, 3, 1, 2, 21, 1, 11, 10)
> x2<-c(26, 29, 56, 31, 52, 55, 71, 31, 54, 47, 40, 66, 68)
> x3<-c(6, 15, 8, 8, 6, 9, 17, 22, 18, 4, 23, 9, 8)
```

表 9.8 水泥数据

序号	X_1	X_2	X_3	X_4	Y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	18	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

```

> x4<-c(60, 52, 20, 57, 33, 22, 6, 44, 22, 18, 34, 12, 12)
> y<-c(78.5, 74.3, 104.3, 87.6, 95.9, 109.2, 102.7, 72.5,
      93.1, 115.9, 83.8, 113.3, 109.4)
> cement<-data.frame(x1, x2, x3, x4)
> lm.reg<-lm(y~x1+x2+x3+x4, data=cement)
> summary(lm.reg)
> library(DAAG)
> vif(lm.reg, digits=3)

```

结果显示为

Call:

```
lm(formula = y ~ x1 + x2 + x3 + x4, data = cement)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
```

```
-3.2777 -1.3956 -0.2374 1.1650 4.0379
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	64.8044	22.8867	2.832	0.02210	*
x1	1.4805	0.3598	4.115	0.00337	**
x2	0.4918	0.2285	2.153	0.06351	.
x3	0.0510	0.3299	0.155	0.88097	
x4	-0.1563	0.2120	-0.737	0.48205	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

因此, 在0.05的水平下, 仅有 X_1 是显著的. 再看一下变量 X_1, X_2, X_3, X_4 的方差膨胀因子

```
> vif(lm.reg, digits=3)
      x1      x2      x3      x4
 9.54 26.90  9.51 31.40
```

结论: 由于 X_2 与 X_4 的方差膨胀因子均大于10, 因此它们之间可能存在共线性. 由命令

```
> cor(x2, x4)
[1] -0.94797
```

知它们之间的线性相关系数达到0.95, 因此可以肯定它们之间的确存在严重的共线性. ■

§9.5 Logistic回归

线性回归模型是定量分析中最常用的统计分析方法, 但线性回归分析要求响应变量是连续型变量. 在实际研究中, 尤其是在生物、医学、经济和社会数据的统计分析中, 研究遇到非连续型的响应变量, 即分类响应变量.

Logistic回归

在研究两元分类响应变量与诸多自变量间的相互关系时,常选用logistic回归模型.

将两元分类响应变量 Y 的一个结果记为“成功”,另一个结果记为“失败”,分别用0和1表示.对响应变量 Y 有影响的 p 个自变量(解释变量)记为 X_1, X_2, \dots, X_p .在 m 个自变量的作用下出现“成功”的条件概率记为 $p = P(Y = 1 | X_1, X_2, \dots, X_p)$,那么logistic回归模型表示为

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p)} \quad (9-5.1)$$

其中 β_0 称为常数项或截距, $\beta_1, \beta_2, \dots, \beta_p$ 称为logistic回归模型的回归系数.

从(9-5.1)式可以看出, logistic回归模型是一个非线性的回归模型, 自变量 $X_j (j = 1, 2, \dots, p)$ 可以是连续变量, 也可以是分类变量, 或哑变量(dummy variable). 对自变量 X_j 任意取值, $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$ 总落在 $(-\infty, +\infty)$ 中, 因此公式(9-5.1)的比值, 即 p 的取值, 总在0到1之间变化, 这是logistic回归模型的合理性所在.

对公式(9-5.1)作logit变换, logistic回归模型可以写成下列线性形式:

$$\text{logit}(p) = \ln \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \quad (9-5.2)$$

这样我们可以使用线性回归模型对参数 $\beta_j, j = 1, 2, \dots, p$ 进行估计.

广义线性模型

logistic回归模型属于广义线性模型(Generalized Linear Model)的一种, 它是通常的正态线性模型的推广, 它要求响应变量只能通过线性形式依赖于解释变量. 上述推广体现在两个方面:

- 通过一个连接函数 ψ , 即对响应变量期望的变换, 将响应变量的期望与解释变量建立线性关系

$$\psi(E(Y)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p.$$

- 通过一个误差函数, 说明广义线性模型的最后一部分随机项;

表9.9给出了广义线性模型中常见的连接函数和误差函数. 可见, 若连接函数为恒等变换, 误差函数为正态分布, 则得到通常的正态线性模型.

表 9.9 常见的连接函数和误差函数

变换	连接函数	回归模型	典型误差函数
恒等	$\psi(x) = x$	$E(y) = X'\underline{\beta}$	正态分布
对数	$\psi(x) = \ln(x)$	$\ln(E(y)) = X'\underline{\beta}$	泊淞分布
logit	$\psi(x) = \text{logit}(x)$	$\text{logit}(E(y)) = X'\underline{\beta}$	二项分布
逆(倒数)	$\psi(x) = 1/x$	$1/E(y) = X'\underline{\beta}$	伽玛分布

与广义线性模型有关的R函数: `glm()`

R语言提供了拟合和计算广义线性模型的函数`glm()`, 其调用格式为

glm() 的调用格式

```
log<-glm(formula, family=family.generator,
          data=data.frame)
```

说明: `formula`为拟合公式, 其意义与线性模型相同; `family`为分布族, 包括正态分布(`gaussian`)、二项分布(`binomial`)、泊淞分布(`poission`)和伽玛分布(`gamma`), 分布族还可通过选项`link`=来指定使用的连接函数; `data`为数据框.

1) 基于正态分布的广义线性模型:

基于正态分布族的glm() 的调用格式

```
log<-glm(formula, family = gaussian(link = identity),
          data = data.frame)
```

说明: `link=identity`可以不写, 因为正态分布族的连接函数默认值是恒等, 再者整个`family=gaussian`也可以不写, 因为分布族的默认值是正态分布. 正态分布族的广义线性模型等同于一般的线性模型, 因此

```
> fm <- glm(formula, family = gaussian, data = data.frame)
```

等同于

```
> fm <- lm(formula, data = data.frame)
```

2) 基于二项分布的广义线性模型:

基于二项分布族的广义线性模型就是本节讲的logistic回归模型, 因此在R软件中, logistic回归分析可以通过调用广义线性回归模型函数glm()来实现, 其调用格式为

—— 基于二项分布族的glm()的调用格式 ——

```
log<-glm(formula, family = binominal(link = logit),  
          data = data.frame)
```

说明: glm()就是R软件中拟合和计算广义线性模型的函数. 公式formula有两种输入方法: 一种是输入成功与失败的次数, 另一种像线性模型通常数据的输入方法. link=logit可以不写, 因为logit是二项分布族连接函数, 是默认状态.

3) 基于泊淞分布的广义线性模型:

—— 基于二项分布族的glm()的调用格式 ——

```
log<-glm(formula, family = poisson(link = log),  
          data = data.frame)
```

4) 基于伽玛分布的广义线性模型:

—— 基于伽玛分布族的glm()的调用格式 ——

```
log<-glm(formula, family = gamma(link = inverse),  
          data = data.frame)
```

例 9.5.1 表9.10为对45名驾驶员的调查结果, 其中四个变量的含义为:

- 1) X_1 : 表示视力状况, 它是一个分类变量, 1表示好, 0表示有问题;
- 2) X_2 : 年龄(age), 数值型;
- 3) X_3 : 驾车(drive)教育, 它也是一个分类变量, 1表示参加过驾车教育, 0表示没有;
- 4) Y : 一个分类型输出变量accident, (去年是否出过事故, 1表示出过事故, 0表示没有).

表 9.10 对45名驾驶员的调查结果

X_1	X_2	X_3	Y	X_1	X_2	X_3	X	X_1	X_2	X_3	Y
1	17	1	1	1	68	1	0	0	17	0	0
1	44	0	0	1	18	1	0	0	45	0	1
1	48	1	0	1	68	0	0	0	44	0	1
1	55	0	0	1	48	1	1	0	67	0	0
1	75	1	1	1	17	0	0	0	55	0	1
0	35	0	1	1	70	1	1	1	61	1	0
0	42	1	1	1	72	1	0	1	19	1	0
0	57	0	0	1	35	0	1	1	69	0	0
0	28	0	1	1	19	1	0	1	23	1	1
0	20	0	1	1	62	1	0	1	19	0	0
0	38	1	0	0	39	1	1	1	72	1	1
0	45	0	1	0	40	1	1	1	74	1	0
0	47	1	1	0	55	0	0	1	31	0	1
0	52	0	0	0	68	0	1	1	16	1	0
0	55	0	1	0	25	1	0	1	61	1	0

试考察前三个变量 X_1, X_2, X_3 与发生事故的关系.

解

1) 用数据框形式输入数据

```
> x1<-rep(c(1, 0, 1, 0, 1), c(5, 10, 10, 10, 10))
> x2<-c(17, 44, 48, 55, 75, 35, 42, 57, 28, 20,
        38, 45, 47, 52, 55, 68, 18, 68, 48, 17,
        70, 72, 35, 19, 62, 39, 40, 55, 68, 25,
        17, 45, 44, 67, 55, 61, 19, 69, 23, 19,
        72, 74, 31, 16, 61),
> x3<-c(1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0,
```

```

1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1,
0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1)
> y<-c(1, 0, 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1,
0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0,
0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0)
> accident<-data.frame(x1, x2, x3, y)

```

2) 再作logistic回归

```

> log.glm<-glm(y~x1+x2+x3, family=binomial, data=accident)
> summary(log.glm)

```

回归结果为:

```

Call: glm(formula = y ~ x1 + x2 + x3, family = binomial,
          data = accident)

```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.5636	-0.9131	-0.7892	0.9637	1.6000

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.597610	0.894831	0.668	0.5042
x1	-1.496084	0.704861	-2.123	0.0338 *
x2	-0.001595	0.016758	-0.095	0.9242
x3	0.315865	0.701093	0.451	0.6523

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

```

Null deviance: 62.183 on 44 degrees of freedom
Residual deviance: 57.026 on 41 degrees of freedom
AIC: 65.026

```

Number of Fisher Scoring iterations: 4

由此得到初步的logistic回归模型:

$$p = \frac{\exp(0.5976 - 1.4961X_1 - 0.0016X_2 + 0.3159X_3)}{1 + \exp(0.5976 - 1.4961X_1 - 0.0016X_2 + 0.3159X_3)}.$$

即

$$\text{logit}(p) = 0.5976 - 1.4961X_1 - 0.0016X_2 + 0.3159X_3.$$

3) 模型的诊断与更新

在此模型中, 由于参数 β_2, β_3 没有通过检验, 可类似于线性模型, 用`step()`做变量筛选.

```
> log.step<-step(log.glm)
> summary(log.step)
```

计算结果为:

```
Start:  AIC= 65.03
y ~ x1 + x2 + x3
```

	Df	Deviance	AIC
- x2	1	57.035	63.035
- x3	1	57.232	63.232
<none>		57.026	65.026
- x1	1	61.936	67.936

```
Step:  AIC= 63.03
y ~ x1 + x3
```

	Df	Deviance	AIC
- x3	1	57.241	61.241
<none>		57.035	63.035
- x1	1	61.991	65.991

```
Step:  AIC= 61.24
y ~ x1
```

```

              Df Deviance    AIC
<none>          57.241 61.241
- x1          1   62.183 64.183

Call: glm(formula = y ~ x1, family = binomial, data = accident)

Deviance Residuals:
      Min       1Q   Median       3Q      Max
-1.4490 -0.8783 -0.8783  0.9282  1.5096

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.6190     0.4688   1.320   0.1867
x1            -1.3728     0.6353  -2.161   0.0307 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 62.183  on 44  degrees of freedom
Residual deviance: 57.241  on 43  degrees of freedom
AIC: 61.241

Number of Fisher Scoring iterations: 4

```

可以看出, 新的回归方程为

$$p = \frac{\exp(0.6190 - 1.3728x_1)}{1 + \exp(0.6190 - 1.3728x_1)}.$$

4) 预测分析:

```

> log.pre<-predict(log.step, data.frame(x1=1))
> p1<-exp(log.pre)/(1+exp(log.pre));p1
> log.pre<-predict(log.step, data.frame(x1=0))
> p2<-exp(log.pre)/(1+exp(log.pre));p2

```

运行得到: $p_1=0.32$; $p_2=0.65$, 说明了视力有问题司机发生交通事故的概率是视力正常的司机的两倍以上.



第九章习题

9.1 测得10名女中学生体重 X_1 (kg)、胸围 X_2 (cm)及肺活量 Y (ml)的数据如表9.11所示, 试画出 Y 与 X_1 , X_2 的散点图, 并分析它们之间的相关关系.

表 9.11 10名女中学生体重 X_1 (kg), 胸围 X_2 (cm)及肺活量 Y (ml)的值

X_1	35	40	40	42	37	45	43	37	44	42
X_2	60	74	64	71	72	68	78	66	70	65
Y	1600	2600	2100	2650	2400	2200	2750	1600	2750	2500

9.2 考察温度对产量的影响, 测得10组数据(见表9.12)

表 9.12 温度对产量的影响

温度 $X/^{\circ}C$	20	25	30	35	40	45	50	55	60	65
产量 Y/kg	13.2	15.1	16.4	17.1	17.9	18.7	19.6	21.2	22.5	24.3

- 1) 试建立 X 与 Y 之间的回归方程式;
- 2) 对其回归方程进行显著性检验;
- 3) 预测 $X = 42^{\circ}C$ 时产量的估计值及预测区间(置信度为95%).

9.3 根据表9.13提供的经济数据,

- 1) 试画出散点图, 判断国民收入(Y)与消费量(X)是否有线性关系;
- 2) 求出 Y 关于 X 的一元线性回归方程;

表 9.13 我国钢材消费量及国民收入

年份	钢材消费量 (万吨)	国民收入 (亿元)	年份	钢材消费量 (万吨)	国民收入 (亿元)
1964	698	1097	1973	1765	2286
1965	872	1284	1974	1762	2311
1966	988	1502	1975	1960	2003
1967	807	1394	1976	1902	2435
1968	738	1303	1977	2013	2625
1969	1025	1555	1978	2446	2948
1970	1316	1917	1979	2736	3155
1971	1539	2051	1980	2825	3372
1972	1561	2111			

- 3) 对方程作显著性检验;
- 4) 现测得1981年消费量 $X = 3441$, 试给出1981年国民收入的预测值及相应的区间估计($\alpha = 0.05$).

9.4 已知变量 X 与 Y 的观测值如表9.14所示.

- 1) 画出数据的散点图, 求回归直线 $Y = \hat{\beta}_0 + \hat{\beta}_1 X$, 同时将回归直线也画在散点图上;
- 2) 对回归模型与参数分别进行 F 检验和 t 检验;
- 3) 画出残差(普通残差和标准残差)与预测值的残差图, 分析误差是否是等方差的;
- 4) 修正模型. 对响应变量 y 作开方, 再完成(1)–(3)的工作.

9.5 某厂生产的一种电器的年销售量 Y 与竞争对手的价格 X_1 及本厂的价格 X_2 有关. 表9.15是10个城市中记录的资料.

- 1) 建立 Y 与 X_1 及 X_2 的回归关系, 并说明回归方程式在 $\alpha = 0.05$ 的水平上是否显著? 并解释回归系数的含义;

表 9.14 数据表

序号	X	Y	序号	X	Y	序号	X	Y
1	1	0.6	11	4	3.5	21	8	17.5
2	1	1.6	12	4	4.4	22	8	13.4
3	1	0.5	13	4	5.1	23	8	4.5
4	1	1.2	14	5	5.7	24	9	30.4
5	2	2.0	15	6	3.4	25	11	12.4
6	2	1.3	16	6	9.7	26	12	13.4
7	2	2.5	17	6	8.6	27	12	26.2
8	3	2.2	18	7	4.0	28	12	7.4
9	3	2.4	19	7	5.5			
10	3	1.2	20	7	10.5			

表 9.15 10个城市某种电器的年销售量和竞争对手价格(单位: 元)

X_1	X_2	Y	X_1	X_2	Y
120	100	102	140	110	100
190	90	120	130	150	77
155	210	46	175	150	93
125	250	26	145	270	69
180	300	65	150	250	85

- 2) 对回归模型进行初步诊断, 并指出有无可疑点或异常点?
- 3) 已知某城市中本厂电器的售价 $X_2 = 160$ 元, 竞争对手售价 $X_1 = 170$ 元, 使用上述建立起来的回归模型预测该城市的年销售量;
- 4) 您能否建立系数 $R^2 > 0.68$, 模型中所有回归系数在0.10水平上是显著的回归模型(考虑二次项和交叉项, 用逐步回归法).

9.6 某科学基金会的管理人员欲了解从事研究的工作人员中, 高水平的数学家工资额 Y 与他们的研究成果(论文、著作等)的质量指标 X_1 , 从事研究工作

的时间 X_2 以及能成功获得资助的指标 X_3 之间的关系, 为此按一定的设计方案调查了24位此类型的数学家, 如表9.16.

表 9.16 24位数学家工资额及相关指标的调查数据

序号	Y	X_1	X_2	X_3	序号	Y	X_1	X_2	X_3
1	33.2	3.5	9	6.1	13	43.3	8.0	23	7.6
2	40.3	5.3	20	6.4	14	44.1	5.6	35	7.0
3	38.7	5.1	18	7.4	15	42.8	6.6	39	5.0
4	46.8	5.8	33	6.7	16	33.6	3.7	31	4.4
5	41.4	4.2	31	7.5	17	34.2	6.2	7	5.5
6	37.5	6.0	13	5.9	18	48.0	7.0	40	7.0
7	39.0	6.8	25	6.0	19	38.0	4.0	35	6.0
8	40.7	5.5	30	4.0	20	35.9	4.5	23	3.5
9	30.1	3.1	5	5.8	21	40.4	5.9	33	4.0
10	52.9	7.2	47	8.3	22	36.8	5.6	27	4.3
11	38.2	4.5	25	5.0	23	45.2	4.8	34	8.0
12	31.8	4.9	11	6.4	24	35.1	3.9	15	5.0

- 1) 假设误差服从 $N(0, \sigma^2)$ 分布, 建立 Y 与 X_1 , X_2 和 X_3 之间的线性回归方程并研究相应的统计推断问题, 作相应的诊断和检验;
- 2) 假定某位数学家的关于 X_1 , X_2 , X_3 的值为 $(x_{01}, x_{02}, x_{03}) = (5.1, 20, 7.2)$, 试预测他的年工资额, 并给出置信度为95%的置信区间.

9.7 某种水泥在凝固时放出的热量 Y (cal/g)与水泥中四种化学成分 X_1 , X_2 , X_3 , X_4 有关, 现测得13组数据, 如表9.17所示.

- 1) 希望从中选出主要变量, 建立 Y 与它们的回归方程;
- 2) 考查 X_1 , X_2 , X_3 , X_4 之间是否存在多重共线性;
- 3) 分析用函数`step()`去掉的变量是否合理.

表 9.17 水泥在凝固时放出的热量与四种化学成分

序号	Y	X ₁	X ₂	X ₃	X ₄	序号	Y	X ₁	X ₂	X ₃	X ₄
1	7	26	6	60	78.5	8	1	31	22	44	72.5
2	1	29	15	52	74.3	9	2	54	18	22	93.1
3	11	56	8	50	104.3	10	21	47	4	26	115.9
4	11	31	8	47	87.6	11	1	40	23	34	83.8
5	7	52	6	33	95.9	12	11	66	9	12	113.3
6	11	55	9	22	109.2	13	10	68	8	12	109.4
7	3	71	17	6	102.7						

表 9.18 两种疗法对不同病情的某病的疗效

	病情	有效(1)	无效(0)
甲药(0)	轻(1)	38	64
	重(0)	10	82
乙药(1)	轻(1)	95	18
	重(0)	50	35
丙药(2)	轻(1)	88	26
	重(0)	43	37

9.8 某研究者欲比较3个不同的药物治疗病情不同的某病的效果, 研究数据见表9.18, 试对数据进行logistic回归分析, 并作相应的统计推断.

9.9 表9.19是40名肺癌病人的生存资料, 其中X₁表示生活行为能力评分(1到100); X₂表示病人的年龄(年); X₃表示由诊断到进入研究时间(月); X₄表示肿瘤类型(“0”是鳞瘤, “1”是小型细胞癌, “2”是腺癌, “3”是大型细胞癌); X₅表示两种化疗方法(“1”是常规, “0”是试验新法); Y表示病人的生存时间(“0”是生存时间短, 即生存时间小于200天; “1”表示生存时间长, 即生存时间大于或等于200天).

表 9.19 40名肺癌病人的生存资料

序号	X_1	X_2	X_3	X_4	X_5	Y	序号	X_1	X_2	X_3	X_4	X_5	Y
1	70	64	5	1	1	1	21	60	37	13	1	1	0
2	60	63	9	1	1	0	22	90	54	12	1	0	1
3	70	65	11	1	1	0	23	50	52	8	1	0	1
4	40	69	10	1	1	0	24	70	50	7	1	0	1
5	40	63	58	1	1	0	25	20	65	21	1	0	0
6	70	48	9	1	1	0	26	80	52	28	1	0	1
7	70	48	11	1	1	0	27	60	70	13	1	0	0
8	80	63	4	2	1	0	28	50	40	13	1	0	0
9	60	63	14	2	1	0	29	70	36	22	2	0	0
10	30	53	4	2	1	0	30	40	44	36	2	0	0
11	80	43	12	2	1	0	31	30	54	9	2	0	0
12	40	55	2	2	1	0	32	30	59	87	2	0	0
13	60	66	25	2	1	1	33	40	69	5	3	0	0
14	40	67	23	2	1	0	34	60	50	22	3	0	0
15	20	61	19	3	1	0	35	80	62	4	3	0	0
16	50	63	4	3	1	0	36	70	68	15	0	0	0
17	50	66	16	0	1	0	37	30	39	4	0	0	0
18	40	68	12	0	1	0	38	60	49	11	0	0	0
19	80	41	12	0	1	1	39	80	64	10	0	0	1
20	70	53	8	0	1	1	40	70	67	18	0	0	1

- 1) 建立 $P(Y=1)$ 对 $X_1 \sim X_5$ 的logistic回归模型, $X_1 \sim X_5$ 对 $P(Y=1)$ 的综合影响是否显著? 哪些变量是主要的影响因素, 显著水平如何? 计算各病人生存时间大于等于200天的概率估计值;
- 2) 用逐步回归法选取自变量, 结果如何? 在所选模型下, 计算病人生存时间大于或等于200天的概率估计值, 并将计算结果与(1)中模型作比较, 差异如何? 哪一个模型更合理?

第十章 多元统计分析介绍

本章概要

- ◇ 主成分分析与因子分析
- ◇ 判别分析
- ◇ 聚类分析
- ◇ 典型相关分析
- ◇ 对应分析

多元统计分析(Multivariable Statistical Analysis)也称多变量统计分析、多因素统计分析或多元分析,是研究客观事物中多变量(多因素或多指标)之间的相互关系和多样品对象之间差异以及以多个变量为代表的多元随机变量之间的依赖和差异的现代统计分析理论和方法.

主成分分析与因子分析的目的是寻找多个变量的“代表”,判别分析能将对象分类到已知类别中,聚类分析按照一定的尺度把对象分类,典型相关分析研究两组变量之间的相关问题,对应分析探究行列变量的关系.

§10.1 主成分分析与因子分析

做衣服时,需要测量人体的许多尺寸,如上体长,手臂长,胸围,颈围,总肩宽等等.然而,这些量之间是否有联系的,能否选出它们的某个线性组合,使之基本能够刻画人对服装的要求.若能,选出的线性组合就是诸多尺寸的主成分或称主分量.

主成分分析(Principle Component Analysis)是把多维空间的相关多变量的数据集,通过降维简化为少量而且相互独立的新综合指标,同时又使简化

后的新综合指标尽可能多的包括原指标群中的主要信息,或是尽可能不损失原有指标的主要信息的一种多元统计分析方法.

为了测验中学生的知识与能力,出40道题目,让若干学生回答,每道题目有一得分,这是可以观测的随机变量,我们希望找出有限个不可观测的潜在变量来解释这40个随机变量.这种分析称为因子分析.这种不可观测的潜在变量一般不能表示为原来随机变量的线性组合,但却是有实际意义的,例如语言表达能力,推理能力,艺术修养能力,历史知识和生活常识等,所以因子分析是寻求潜在变量的一种方法.

因子分析(Factor Analysis)最早于1904年由英国著名统计学家、心理学家查尔斯·皮尔逊(Charles.S.Pearson)提出,主要目的是研究相关矩阵的内在依赖关系,把多个显在的变量综合为少数几个不可观测的“潜在因子”或称公共因子,来说明复杂多变量系统的内部结构,并解释原始显在复杂多变量与少数“潜在因子”之间的内在联系和相关关系.然后,根据专业知识和定性分析对综合因子所反映的独特含义进行命名和解释的一种多元统计分析方法.

10.1.1 主成分的简要定义与计算

定义 10.1.1 设 $X = (x_1, x_2, \dots, x_p)'$ 是 p 维随机向量,二阶矩存在,若向量 $t'_1 = (t_{11}^*, t_{12}^*, \dots, t_{1p}^*)$ 在条件 $|t_1| = 1$ 下使得 $\text{Var}(t'_1 X)$ 最大,则称 $Y_1 = t'_1 X$ 是 X 的第一主成分或第一主分量;若向量 $t'_2 = (t_{21}^*, t_{22}^*, \dots, t_{2p}^*)$ 在条件 $|t_2| = 1, \text{Cov}(t'_2 X, Y_1) = 0$ 下使得 $\text{Var}(t'_2 X)$ 最大,则称 $Y_2 = t'_2 X$ 是 X 的第二主成分或第二主分量等等.

由定义可见, Y_1 尽可能多地反映原来 p 个变量的信息, Y_2 在与 Y_1 不相关条件下尽可能多地反映原来 p 个变量的信息,这样继续下去.定理10.1.1给出主成分的计算公式.

定理 10.1.1 设 X 为 p 维随机向量, $\text{Cov}(X) = \Sigma$ 存在,则 X 的第 i 个主成分为 $Y_i = t'_i X, i = 1, 2, \dots, p$, 其中 $\text{Var}(Y_i) = \lambda_i$ 是 Σ 的特征值从大到小排序后第 i 个特征值; t_i 是 λ_i 的特征向量.

定义 10.1.2 $\lambda_k / \sum_{i=1}^p \lambda_i$ 称为主成分 Y_k 的方差贡献率; $\sum_{i=1}^k \lambda_i / \sum_{i=1}^p \lambda_i$ 称为主成分 Y_1, Y_2, \dots, Y_k 的累计方差贡献率, Y_k 与 X 第 i 个分量的相关系数 $\rho(x_i, Y_k)$ 称为因子载荷量.

易证明 $\rho(x_i, Y_k) = \sqrt{\lambda_k} t_{ki} / \sigma_i$, 其中 σ_i^2 是 x_i 的方差, t_{ki} 是 t_k 第 i 个分量.

通常取 m 使 Y_1, Y_2, \dots, Y_m 的累计方差贡献率达到70%或80%以上, 然后考虑用 Y_1, Y_2, \dots, Y_m 来描述 X 的性质.

在实际问题中, X 的不同分量有时有不同的量纲, 量纲变小时该分量的方差会变大, 从而在主成分中变得突出, 造成不合理的结果. 为了避免量纲的影响, 常常将随机变量都标准化, 令

$$x_i^* = \frac{x_i - E(x_i)}{\sqrt{\text{Var}(x_i)}}, \quad i = 1, 2, \dots, p \quad (10-1.1)$$

$X^* = (x_1^*, x_2^*, \dots, x_p^*)'$, 再来求 X^* 的主成分, 而 X^* 的协差阵就是 X^* 的相关阵, 也是 X 的相关阵 R , 因此我们如下的定理.

定理 10.1.2 设 X 的相关阵为 R , 其特征值 $\lambda_1^* \geq \lambda_2^* \geq \dots \geq \lambda_p^*$, 相应特征向量为 $t_1^*, t_2^*, \dots, t_p^*$, 则 X^* 的主成分分别是 $Y_1^* = t_1^{*'} X^*, Y_2^* = t_2^{*'} X^*, \dots, Y_p^* = t_p^{*'} X^*$. x_i^* 与主成分 Y_k^* 的相关系数(因子载荷量)为 $\rho(x_i^*, Y_k^*) = \sqrt{\lambda_k^*} t_{ki}^*$, 其中 t_{ki}^* 是 t_k^* 的第 i 个分量.

实际问题中协差阵、相关阵都是未知的, 总用样本协差阵与样本相关阵代替, 这样是有道理的: 若 $X \sim N(\mu, \Sigma)$, $\hat{\Sigma}$ 是 Σ 的极大似然估计, $\hat{\Sigma}$ 的特征值为 $\nu_1 \geq \nu_2 \geq \dots \geq \nu_p$, 相应单位特征向量 $\iota_1, \iota_2, \dots, \iota_p$; 而 Σ 的特征值, 特征向量为 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p, t_1, t_2, \dots, t_p$. 则可以证明

定理 10.1.3 $\nu_1, \nu_2, \dots, \nu_p$ 是 $\lambda_1, \lambda_2, \dots, \lambda_p$ 的极大似然估计, $\iota_1, \iota_2, \dots, \iota_p$ 是 t_1, t_2, \dots, t_p 的极大似然估计.

10.1.2 主成分R通用程序

利用R语言的`princomp()`函数就可完成主成分分析, `princomp()`的二种调用格式如下:

—— `princomp()` 的调用格式-1 ——

```
princomp(formula, data = NULL, subset, na.action, ...)
```

或者

—— `princomp()` 的调用格式-2 ——

```
princomp(x, cor = FALSE, scores = TRUE, covmat = NULL,
```

```
subset = rep(TRUE, nrow(as.matrix(x))), ...)
```

说明: `formula`是没有响应变量的公式; `x`是用于主成分分析的数据; `cor`是逻辑变量, 当`cor=TRUE`表示用样本的相关阵 \mathbf{R} 作主成分分析, 否则当`cor=FALSE`(默认选项)表示用样本的协方差阵 S 作主成分, 具体说明见 \mathbf{R} 帮助.

例 10.1.1 (学生身体4项指标的主成分分析) 随机抽取30名某年级中学生, 测量其身高(X_1), 体重(X_2), 胸围(X_3), 坐高(X_4), 数据如下表所示, 试对这30名学生身体四项指标作主成分分析.

表 10.1 30名学生的4项指标

序号	X_1	X_2	X_3	X_4	序号	X_1	X_2	X_3	X_4
1	148	41	72	78	2	139	34	71	76
3	160	49	77	86	4	149	36	67	79
5	159	45	80	86	6	142	31	66	76
7	153	43	76	83	8	150	43	77	79
9	151	42	77	80	10	139	31	68	74
11	140	29	64	74	12	161	47	78	84
13	158	49	78	83	14	140	33	67	77
15	137	31	66	73	16	152	35	73	79
17	149	47	82	79	18	145	35	70	77
19	160	47	74	87	20	156	44	78	85
21	151	42	73	82	22	147	38	73	78
23	157	39	68	80	24	147	30	65	75
25	157	48	80	88	26	151	36	74	80
27	144	36	68	76	28	141	30	67	76
29	139	32	68	73	30	148	38	70	78

解 \mathbf{R} 程序如下:

```

> student<-data.frame(
  X1=c(148, 139, 160, 149, 159, 142, 153, 150, 151, 139,
       140, 161, 158, 140, 137, 152, 149, 145, 160, 156,
       151, 147, 157, 147, 157, 151, 144, 141, 139, 148),
  X2=c(41, 34, 49, 36, 45, 31, 43, 43, 42, 31,
       29, 47, 49, 33, 31, 35, 47, 35, 47, 44,
       42, 38, 39, 30, 48, 36, 36, 30, 32, 38),
  X3=c(72, 71, 77, 67, 80, 66, 76, 77, 77, 68,
       64, 78, 78, 67, 66, 73, 82, 70, 74, 78,
       73, 73, 68, 65, 80, 74, 68, 67, 68, 70),
  X4=c(78, 76, 86, 79, 86, 76, 83, 79, 80, 74,
       74, 84, 83, 77, 73, 79, 79, 77, 87, 85,
       82, 78, 80, 75, 88, 80, 76, 76, 73, 78)
)
> student.pr<-princomp(student, cor=TRUE)
> summary(student.pr,loadings=TRUE)

```

计算结果为:

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	1.8817805	0.55980636	0.28179594	0.25711844
Proportion of Variance	0.8852745	0.07834579	0.01985224	0.01652747
Cumulative Proportion	0.8852745	0.96362029	0.98347253	1.00000000

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4
X1	-0.497	0.543	-0.450	0.506
X2	-0.515	-0.210	-0.462	-0.691
X3	-0.481	-0.725	0.175	0.461
X4	-0.507	0.368	0.744	-0.232

对上述结果我们作一些说明:

- 1) **Standard deviation:** 表示主成分的标准差, 即主成分的方差平方根, 即相应特征值的开方;

- 2) Proportion of Variance: 表示方差的贡献率;
- 3) Cumulative Proportion: 表示方差的累计贡献率.
- 4) 用summary函数中loadings=TRUE选项列出了主成分对应原始变量的系数, 因此得到前两个主成分是
- $$Y_1 = -0.497x_1^* + 0.543x_2^* - 0.450x_3^* + 0.506x_4^*$$
- $$Y_2 = -0.515x_1^* - 0.210x_2^* - 0.462x_3^* + 0.691x_4^*$$
- 由于前两个主成分的累计贡献率已经达到96.36%, 所以取前两个主成分来降维.
- 5) 对于主成分的解释: 由 Y_1 的系数都接近与0.5, 它反映学生身材的魁梧程度, 因此我们称第一主成分为大小因子(魁梧因子); Y_2 的系数中体重(X_2)和胸围(X_3)为正值, 它反映学生的胖瘦情况, 故称第二主成分为形状因子(或胖瘦因子).

■

10.1.3 因子分析的简要定义与计算

因子分析方法根据研究对象和分析方法的不同, 分为R型和Q型两种不同的类型. R型因子分析研究指标(变量)之间的相互关系, 通过对多变量相关系数矩阵内部结构的研究, 找出控制所有变量的几个主因子(主成分); Q型因子分析研究样品之间控制所有样品的几个主要因素. 由于这两种因子分析方法的相关关系, 所以通过样品相似系数矩阵与通过变量相关系数矩阵内部结构的研究, 找出分析的全部运算过程都是一样的, 只是出发点不同而已. R型分析从相关系数矩阵出发, Q型分析从相似系数矩阵出发, 对于同一批观测数据, 可根据所要求的目的决定采用哪一类型的分析. 只是R型分析须考虑变量量纲及数量级, 而Q型分析则不必考虑这一问题, 在多变量的量纲及数量级差别很大时, 更为方便. 而对于同一批观测数据, 可以根据其所要求的目的而决定采用哪一类型的分析.

定义 10.1.3 设 X 为 $p \times 1$ 随机向量, 其均值为 μ , 协差阵为 $\Sigma = (\sigma_{ij})$, 若 X 能表示为

$$X = \mu + \Lambda f + u \quad (10-1.2)$$

其中 Σ 是 $p \times k$ 未知常数阵, f 是 $k \times 1$ 随机变量, μ 是 $p \times 1$ 随机向量, 且

$$\begin{cases} E(f) = 0, \text{Var}(f) = I \\ E(\mu) = 0, \text{Var}(\mu) = \Psi = \text{diag}(\Psi_1^2, \Psi_2^2, \dots, \Psi_p^2) \\ \text{Cov}(f, \mu) = 0 \end{cases} \quad (10-1.3)$$

则 $X = \mu + \Lambda f + u$ 称为 X 有 k 个因子的因子分析模型, f 称为公共因子, μ 称为特殊因子, Σ 叫做因子载荷矩阵, 其元素 δ_{ij} 是第 i 个变量在第 j 个因子上的载荷.

由上面的关系我们可见 $\text{Cov}(X) = \Lambda\Lambda' + \Psi = \Sigma$, 从而 Σ 对角线上元素

$$\sigma_{ii} = \sum_{j=1}^k \lambda_{ij}^2 + \Psi_i^2 = h_i^2 + \Psi_i^2, i = 1, 2, \dots, p \quad (10-1.4)$$

其中 h_i^2 反映了公共因子对 X_i 的影响, 称为共同度或共性方差.

值得注意的是, 因子载荷不是唯一的, 若 Γ 是任意 k 阶正交阵, 则 X 可以表示成 $X = \mu + (\Lambda\Gamma)(\Gamma'f) + u$. 将 $\Lambda\Gamma$ 作为因子载荷, $\Gamma'f$ 作为公共因子, 则(3)式仍然成立, 因子载荷的不唯一性, 使得我们有更多地选择余地, 反而是有利的.

实际问题中, 总是给出随机向量的 n 个观测值, 从而得到样本方差阵, 进而估计因子载荷, 并给公因子赋予有实际背景的解释.

模型 $X = \mu + \Lambda f + u$ 中 μ 可用样本均值来估计. Σ 可用 $\sum_{i=1}^n (X^{(i)} - \bar{X})(X^{(i)} - \bar{X})' / (n-1)$ 估计, 其中 $X^{(i)}$ 是随机向量的第 i 次观察值.

提取因子的方法有多种, 常用的有主成分分析、主因子分析、迭代主因子分析、极大似然分析等, 用上述方法之一估计出参数后, 还必须对得到的公共因子进行解释, 对每个公共因子要给出一个名称, 说明其作用. 有时公共因子 f 难以和实际问题相对应, 这时需要通过某个正交阵 Γ 作公共因子旋转, 使 $\Gamma'f$ 和 $\Lambda\Gamma$ 有鲜明的实际意义. 另外一方面, 上述方法估计参数带有随意性, 通过旋转公因子, 可以减少随意性. 所以作公共因子旋转是有必要的. 有最大方差旋转、最大均方旋转等旋转方法.

因子分析与主成分分析的形式上类似, 但有着明显的区别, 主要表现在五个方面:

- 1) 因子分析需要构造因子模型, 是把原观测变量表现为公共因子(新综合因子)与特殊因子的有机组合模型. 而主成分分析不能作为一个模型来描

述, 只能作为通常的变量变换, 也就是把新综合变量表现为原多变量的线性变换(组合);

- 2) 在理论上主成分分析中的综合主分量数 m 和原变量的个数 p 之间是相等的, 它是把一组具有相关性的变量变换为一组新的独立变量. 而因子分析的目的是要求构造的因子模型中公共因子的数目尽可能少, 以便尽可能构造一个结构简单的模型;
- 3) 因子分析是把原观测变量表示为新综合因子的线性组合, 即新因子的综合指标, 而主成分分析是把主分量表示为原观测变量的线性组合. 另外, 因子分析模型在形式上与线性回归模型相似, 但两者之间有本质的区别: 回归模型中的自变量是可观测量, 而因子模型中各个公共因子是不可观测的潜在因子, 而且两个模型的参数意义上很不相同;
- 4) 主成分分析的数学模型实质上是一种变换, 而因子分析模型是描述原指标 X 协方差阵 Σ 结构的一种模型;
- 5) 在主成分分析中每个主成分相应的系数是唯一确定的, 而在因子分析中每个因子的相应系数不是唯一的, 即因子载荷阵不是唯一的.

10.1.4 因子分析R通用程序

利用R语言的factanal()函数就可完成因子分析, 其基本的调用格式如下:

factanal() 的调用格式

```
factanal(x, factors, data = NULL, covmat = NULL, n.obs = NA,
        subset, na.action, start = NULL,
        scores = c("none", "regression", "Bartlett"),
        rotation = "varimax", control = NULL, ...)
```

说明: x 是用于因子分析的数据; $factors$ 表示因子个数, $scores$ 表示选用因子得分的方法, $rotation = "varimax"$ 表示用最大方差旋转, 具体说明见R帮助.

例 10.1.2 100名学生六门课程(数学、物理、化学、语文、历史、英语)的成绩如下表(只列出了部分, 数据在student.txt). 目前的问题是, 能不能把这个数据的6个变量用一两个综合变量来表示呢? 这一两个综合变量包含有多少原来的信息呢? 怎么解释它们呢?

表 10.2 100名学生六门课程成绩

学生代码	数学	物理	化学	语文	历史	英语
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
4	80	69	75	74	74	63
5	74	70	80	84	81	74
6	78	84	75	62	71	64
7	66	71	67	52	65	57
8	77	71	57	72	86	71
9	83	100	79	41	67	50
...

解 R程序如下:

```
> student<-read.table("D:/Rdata/student.txt")
> names(student)=c("math", "phi", "chem", "lit", "his", "eng")
> fa<-factanal(student, factors=2)
> fa
```

R程序结果:

```
Call: factanal(x = student, factors = 2)
```

Uniquenesses:

```
math phi chem lit his eng
0.245 0.451 0.479 0.136 0.215 0.181
```

Loadings:

```
Factor1 Factor2
math -0.355 0.793
phi -0.201 0.713
```

```
chem -0.216    0.689
lit   0.850   -0.376
his   0.854   -0.235
eng   0.872   -0.242
```

```

                Factor1 Factor2
SS loadings      2.425    1.868
Proportion Var   0.404    0.311
Cumulative Var   0.404    0.716
```

Test of the hypothesis that 2 factors are sufficient.
 The chi square statistic is 0.39 on 4 degrees of freedom.
 The p-value is 0.983

结果说明:

- 1) 我们用 $x_1, x_2, x_3, x_4, x_5, x_6$ 来表示math(数学), phys(物理), chem(化学), literat(语文), history(历史), english(英语)等变量. 这样因子 f_1 和 f_2 与这些原变量之间的关系是

$$\begin{aligned}
 x_1 &= -0.355f_1 + 0.793f_2 \\
 x_2 &= -0.201f_1 + 0.713f_2 \\
 x_3 &= -0.216f_1 + 0.689f_2 \\
 x_4 &= 0.850f_1 - 0.376f_2 \\
 x_5 &= 0.854f_1 - 0.235f_2 \\
 x_6 &= 0.872f_1 - 0.242f_2
 \end{aligned} \tag{10-1.5}$$

这里, 第一个因子主要和语文、历史、英语三科有很强的正相关, 相关系数分别为0.850, 0.854, 0.872; 而第二个因子主要和数学、物理、化学三科有很强的正相关相关系数分别为0.793, 0.713, 0.689. 因此可以给第一个因子起名为“文科因子”, 而给第二个因子起名为“理科因子”.

- 2) Proportion Var 是方差贡献率, Cumulative Var是累计方差贡献率, 检验表明两个因子已经充分.

■

§10.2 判别分析

判别分析是用于判断样品所属类型的一种统计分析方法. 判别分析的目的在于对已知归类的数据建立由数值指标构成的归类规则, 然后把这样的规则应用到未知归类的样品去归类. 在生产、科研和日常生活中经常会遇到如何根据观测到的数据资料对所研究的对象进行判别归类的问题. 例如一个病人肺部有阴影, 医生需要判断他患的是肺结核、肺部良性肿瘤还是肺癌. 这里, 肺结核病人, 肺部良性肿瘤病人和肺癌病人组成了三个总体, 病人可能就来源于这三个总体之一, 判别分析的目的在于通过病人的指标(阴影大小, 阴影部位, 边缘是否光滑, 是否有痰, 是否有热度…)来判断他该属于哪个总体(即判别他生的是什么病). 又如根据已有的气象资料(气温, 气压等)来判断明天是晴天还是阴天, 是有雨还是无雨, 所以判别分析是应用性很强的一种多元分析的方法.

判别分析的一般提法是: 设有 k 个总体 G_1, G_2, \dots, G_k , 已知样品 X 来自这 k 个总体的某一个, 但不知它究竟来自哪一个. 判别分析就是要根据对这 k 个总体的已知知识(由过去的经验或抽样获得)和待判样品的一些指标的观测值, 去判别样品 X 应归属于哪一个总体.

如同经典的数理统计分析, 我们对于这 k 个总体 G_1, G_2, \dots, G_k 的了解程度在不同的场合不尽相同. 有时其分布函数完全已知, 设为 $F_1(x), F_2(x), \dots, F_k(x)$; 有时只知道其形式, 其中某个或某些未知参数未知; 有时我们对于它们全然不知. 前面二种场合下的判别分析称为参数判别方法, 后面一种场合下的判别分析称为非参数判别方法.

通常我们先对预先得到的来自这 k 个总体的若干个样品(称为训练样品)进行检验和归类, 来决定相应的判别归类问题是否有意义及误判可能性大小. 然后再对给定的一个或几个新的样品, 进行判别归类, 即决定它(们)自哪个总体. 解决这个问题可以有多种途径, 下面我们分别讨论几种常用的方法, 如距离判别、Fisher判别等.

10.2.1 距离判别

距离判别法(或称直观判别法)的基本思想是: 样品和哪个总体距离最近, 就判它属于哪个总体.

两个总体的距离判别

设有两个总体(或称两类) G_1, G_2 , 从第一个总体中抽取 n_1 个样品, 从第二

个总体中抽取 n_2 个样品, 每个样品观测 m 个指标 x_1, \dots, x_m . 所得的数据集称为训练样本

今取一个样品 X , 实测指标值 $X = (x_1, \dots, x_m)$, 问该样品应该判为哪一类?

首先计算样品 X 到 G_1 和 G_2 两类的距离, 分别记为 $D(X, G_1)$ 和 $D(X, G_2)$, 按照距离判别归类, 即: 样品离哪个总体距离最近, 就判它属于哪个总体; 如果样品到两个总体距离相等, 则暂时不归类. 判别准则可以写为:

$$\begin{cases} X \in G_1, \text{ 如果 } D(X, G_1) < D(X, G_2) \\ X \in G_2, \text{ 如果 } D(X, G_2) < D(X, G_1) \\ X \text{ 待判, 如果 } D(X, G_1) = D(X, G_2) \end{cases} \quad (10-2.1)$$

距离 D 的定义有很多种, 但是考虑到判别分析中常涉及多个变量的问题, 且变量之间可能有相关性, 故多用马氏(Mahalanobis)距离:

$$D(X, G) = (X - \mu)' \Sigma^{-1} (X - \mu) \quad (10-2.2)$$

其中 $\mu = (\mu_1, \dots, \mu_m)'$ 为 G 的均值向量, $\Sigma = (\sigma_{ij})_{m \times m}$ 为 G 的协方差阵.

在实际问题中, 通常 $G_i (i=1, 2)$ 的均值向量 μ_i 和协方差阵 Σ_i 均未知, 故需要由来自它们的训练样品 $X_t^{(i)}, t = 1, 2, \dots, n_i, i = 1, 2$ 进行估计. 它们的极大似然估计分别为

$$\bar{X}^{(i)} = \frac{1}{n_i} \sum_{t=1}^{n_i} X_t^{(i)}, i = 1, 2$$

$$S_i = \frac{1}{n_i - 1} \sum_{t=1}^{n_i} (X_t^{(i)} - \bar{X}^{(i)})(X_t^{(i)} - \bar{X}^{(i)})', i = 1, 2$$

特别地, 若假定两总体的协方差阵相等, 则它们的共同的协方差阵 $\Sigma = \Sigma_1 = \Sigma_2$ 就用它们的样本合并协方差阵 S 进行估计:

$$S = \frac{1}{n-2} [(n_1 - 1)S_1 + (n_2 - 1)S_2], \quad n = n_1 + n_2.$$

这时可由两马氏距离之差得到线性判别函数 $W(X) = a'(X - X^*)$, 其中

$$a = S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}), X^* = (\bar{X}^{(1)} + \bar{X}^{(2)})/2.$$

相应的判别规则变成

$$\begin{cases} X \in G_1, \text{如果} W(X) > 0 \\ X \in G_2, \text{如果} W(X) < 0 \\ X \text{待判, 如果} W(X) = 0 \end{cases} \quad (10-2.3)$$

多个总体的距离判别

类似与两个总体的情况, 多个总体的情况, 按照距离最近的原则对 X 进行判别归类时, 首先计算样品到各类的马氏(Mahalanobis)距离, 然后进行比较, 把待判样品判归距离最小的那个总体. 计算马氏(Mahalanobis)距离时, 类似地可以考虑 $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$ 或者 Σ_i 不仅相等的两种情况.

这种根据距离远近来判别的方法, 原理简单, 直观易懂, 且是后面要介绍的Fisher判别的基础.

10.2.2 Fisher判别法

Fisher判别的基本思想是投影. 将 k 组 m 维数据投影到某个方向, 使得投影后组与组之间尽可能地分开. 而衡量组与组之间是否分开的方法借助于一元方差分析的思想.

设从 p 维总体 $G_t (t = 1, 2, \cdots, k)$ 中分别抽取 n_t 个样品 $X_j^{(t)}, j = 1, 2, \cdots, n_t$, 令 $a = (a_1, a_2, \cdots, a_p)'$ 为 p 维空间中的任一向量, $u(X) = a'X$ 表示 X 向以 a 为法线方向的投影. 通过这样的投影, 可以将原来的数据转化为 k 组一维数据: $a'X_j^{(t)}, j = 1, 2, \cdots, n_t, t = 1, 2, \cdots, k$. 按一元方差分析的思想, 其组间平方和为

$$\begin{aligned} B_0 &= \sum_{t=1}^k n_t (a' \bar{X}_j^{(t)} - \bar{X})^2 \\ &= a' \left[\sum_{t=1}^k n_t (\bar{X}_j^{(t)} - \bar{X})(\bar{X}_j^{(t)} - \bar{X})' \right] a = a' B a. \end{aligned}$$

合并的组内平方和为

$$E_0 = a' \left[\sum_{t=1}^k \sum_{j=1}^{n_i} (\bar{X}_j^{(t)} - \bar{X})(\bar{X}_j^{(t)} - \bar{X})' \right] a = a' E a,$$

其中 $\bar{X}_j^{(t)}$ 和 \bar{X} 分别为 G_t 的样本均值和总样本均值. 若 k 类的均值有显著差异, 则比值 $\Delta(a) = \frac{a'Ba}{a'Ea}$. 应该充分大. 利用方差分析的思想, 问题化为求投影方向 a , 使得 $\Delta(a)$ 达到极大值, 但 $\Delta(a)$ 达到极大值的 a 并不唯一. 等价地, 我们可以对 a 加一约束条件, 即选取 a 使得 $a'Ea = 1$. 问题化为求 a , 使 $\Delta(a) = a'Ba$ 在 $a'Ea = 1$ 条件下达极大.

利用Lagrange乘数法可以容易地导出线性判别函数 $u(X) = a'X$, 其中 a 为特征方程 $|E^{-1}B - \lambda I| = 0$ 的最大特征根所对应的满足 $a'Ea = 1$ 的特征向量.

若仅用一个线性判别函数不能很好地区分各个总体, 则可用第二大特征根、第三大特征根...对应的特征向量构造线性判别函数进行判别, 线性判别函数的个数不超过 $k-1$ 个. 判别的效率用这些特征根来度量.

10.2.3 R通用程序

首先我们要用命令

```
>library(MASS)
```

加载MASS宏包, 再用函数 `lda()` 就可完成Fisher判别分析, 其基本调用格式如下:

lda() 的调用格式

```
lda(formula, data, ..., subset, na.action)
```

说明: `formula`用法为 `groups ~ x1 + x2 + ...`, `group`表明总体来源, `x1, x2, ...`表示分类指标; `subset`指明训练样本. 具体说明见R帮助.

例 10.2.1 Fisher于1936年发表的鸢尾花(Iris)数据被广泛地作为判别分析的例子. 数据是对3品种(species)鸢尾花: 刚毛鸢尾花(setosa)、变色鸢尾花(versicolor)、弗吉尼亚鸢尾花(virginica)各抽取一个容量为50的样本, 测量其花萼长(Sepal.Length)、花萼宽(Sepal.Width)、花瓣长(Petal.Length)、花瓣宽(Petal.Width), 单位为mm. 试调用R内置档案中的iris数据文件进行判别分析.

解 R程序如下:

```
> data(iris)
> attach(iris)
```

```

> names(iris)
> library(MASS)
> iris.lda <- lda(Species ~ Sepal.Length + Sepal.Width
                  + Petal.Length + Petal.Width)
> iris.lda
> iris.pred=predict(iris.lda) $ class
> table(iris.pred, Species)
> detach(iris)

```

`predict()` 是R内置函数, 可以将`lda()`的输出应用于原本iris的数据进行预测, 从而进行对比.

R程序结果:

```

Call: lda(Species ~ Sepal.Length + Sepal.Width
          + Petal.Length +Petal.Width)

```

Prior probabilities of groups:

```

      setosa versicolor virginica
0.3333333  0.3333333  0.3333333

```

Group means:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
setosa	5.006	3.428	1.462	0.246
versicolor	5.936	2.770	4.260	1.326
virginica	6.588	2.974	5.552	2.026

Coefficients of linear discriminants:

	LD1	LD2
Sepal.Length	0.8293776	0.02410215
Sepal.Width	1.5344731	2.16452123
Petal.Length	-2.2012117	-0.93192121
Petal.Width	-2.8104603	2.83918785

Proportion of trace:

```

LD1    LD2

```

```
0.9912 0.0088
Species
iris.pred  setosa  versicolor virginica
setosa      50      0      0
versicolor  0      48      1
virginica   0      2      49
```

结果说明:

- 1) Group means: 包含了每组的平均向量
- 2) Coefficients of linear discriminants: 线性判别系数
- 3) Proportion of trace: 表明了第*i*判别式对区分各组的贡献大小
- 4) Species: 表明将原始数据代入线性判别函数后的判别结果, **setosa**组没有错判, **versicolor**有两个错判, **virginica**只有一个错判.

■

例 10.2.2 盐泉含钾性判别(见表10.3, 并生成数据文件disc.txt): 某地区经勘探证明A盆地是一个钾盐矿区, B盆地是一个钠盐矿区, 其他盐盆地是否含钾盐有待作出判断. 今从A, B两盆地各抽取5个盐泉样品; 从其他盆地抽得8个盐泉样品, 18个盐泉的四个指标数值见下表. 试对后8个待判盐泉进行含钾性判别.

解 R程序如下:

```
> w <- read.table("D:/Rdata/disc.txt")
> names(w)=c("group", "x1", "x2", "x3", "x4")
> library(MASS)
> z <- lda(group~x1+x2+x3+x4, data=w, prior=c(1, 1)/2)
> newdata<-rbind(
      c(8.85, 3.38, 5.17, 26.10), c(28.60, 2.40, 1.20, 127.0),
      c(20.70, 6.70, 7.60, 30.20), c(7.90, 2.40, 4.30, 33.20),
      c(3.19, 3.20, 1.43, 9.90), c(12.40, 5.10, 4.43, 24.60),
      c(16.80, 3.40, 2.31, 31.30), c(15.00, 2.70, 5.02, 64.00))
> dimnames(newdata)<-list(NULL, c("x1", "x2", "x3", "x4"))
> newdata<-data.frame(newdata)
```


表 10.3 盐泉含钾数据

盐泉类别	序号	X_1	X_2	X_3	X_4	类别号
第一类: 含钾盐泉 (A盆地)	1	13.85	2.79	7.80	49.60	A
	2	22.31	4.67	12.31	47.80	A
	3	28.82	4.63	16.18	62.15	A
	4	15.29	3.54	7.50	43.20	A
	5	28.79	4.90	16.12	58.10	A
第二类: 不含钾 盐泉 (B盆地)	6	2.18	1.06	1.22	20.60	B
	7	3.85	0.80	4.06	47.10	B
	8	11.40	0.00	3.50	0.00	B
	9	3.66	2.42	2.14	15.10	B
	10	12.10	0.00	5.68	0.00	B
待 判 盐 泉	1	8.85	3.38	5.17	26.10	
	2	28.60	2.40	1.20	127.0	
	3	20.70	6.70	7.60	30.20	
	4	7.90	2.40	4.30	33.20	
	5	3.19	3.20	1.43	9.90	
	6	12.40	5.10	4.43	24.60	
	7	16.80	3.40	2.31	31.30	
	8	15.00	2.70	5.02	64.00	

```
> predict(z, newdata=newdata)
```

R程序结果:

```
$class [1] B A A B B A A A Levels: A B
```

```
$posterior
```

```
          A          B
```

```
1 1.639701e-03 9.983603e-01
```

```

2 1.000000e+00 1.932625e-83
3 1.000000e+00 1.269619e-20
4 8.302424e-02 9.169758e-01
5 1.190922e-06 9.999988e-01
6 1.000000e+00 1.129611e-10
7 1.000000e+00 1.161894e-26
8 1.000000e+00 7.135903e-22

```

\$x

```

          LD1
1    1.0536512
2   -31.2985593
3    -7.5286829
4    0.3947245
5    2.2416596
6   -3.7639282
7   -9.8136273
8   -8.0017623

```

结果说明:

- 1) 由\$*class*可以看出8个待判样品, 待判样品1, 4, 5属于含钾盐泉(A盆地), 其余属于不含钾盐泉(B盆地);
- 2) \$*x*给出了线性判别函数的数值.

■

§10.3 聚类分析

聚类分析(cluster analysis)是研究“物以类聚”的一种方法, 在国内曾有人称它为群分析、点群分析、簇群分析等. 人类认识世界往往首先将被认识的对象进行分类, 因此分类学便成了人类认识世界的基础科学. 在古老的分类学中, 人们主要靠经验和专业知识实现分类. 随着人类对自然的认识不断加深, 分类越来越细, 要求越来越高, 以致有时光凭借经验和专业知识还不能进行确切的分类, 于是数学这个有用的工具逐渐被引进到分类学中, 形成了数值分类学. 后

来随着多元分析的引进,从数值分类学中逐渐地分离出了聚类分析这个分支.和多元分析的其他方法相比,聚类分析的方法是比较粗糙的,理论尚不完善,但由于它的应用取得很大的成功,和回归分析、判别分析一起被称为多元分析的三大方法.

值得一提的是聚类分析和判别分析都是研究分类问题,但两者有本质的区别.聚类分析一般是寻求客观分类的方法,事先对总体到底有几种类型从无知晓,而判别分析则是在总体类型划分已知,在各总体分布或来自各个总体训练样本的基础上,对当前的新样品用统计分析的方法判定它们属于哪个总体.

10.3.1 基本思想

系统聚类法是将 n 个样品分成若干类的方法,其基本思想是:先将 n 个样品各自看成一类,然后规定类与类之间的距离(类之间的距离有多种定义方法),选择距离最小的一对合并成新的类,计算新类与其他类的距离,再将距离最近的两类合并,这样每次减少一类,直至所有的样品都成为一类为止.

对于距离常用的有以下几种:

(1) 绝对值距离(**R**语言中用**Manhattan**表示),用公式表示为

$$d_{ij}(1) = \sum_{k=1}^p |x_{ik} - x_{jk}| \quad (10-3.1)$$

(2) 欧氏距离(**Euclidean**),用公式表示为

$$d_{ij}(2) = \left[\sum_{k=1}^p (x_{ik} - x_{jk})^2 \right]^{\frac{1}{2}} \quad (10-3.2)$$

(3) 明考斯基距离(**Minkowski**),用公式表示为

$$d_{ij}(q) = \left[\sum_{k=1}^p (x_{ik} - x_{jk})^q \right]^{\frac{1}{q}}, (p > 0) \quad (10-3.3)$$

(4) 切贝雪夫距离(**R**语言中用**maximum**表示),用公式表示为

$$d_{ij}(\infty) = \max_{1 \leq k \leq p} |x_{ik} - x_{jk}| \quad (10-3.4)$$

(5) 马氏距离, 用公式表示为

$$d_{ij}(M) = (X_{(i)} - X_{(j)})' S^{-1} (X_{(i)} - X_{(j)}) \quad (10-3.5)$$

式中 S 是样本协方差矩阵

(6) 兰氏距离(**R**语言中用**canberra**表示), 用公式表示为

$$d_{ij}(L) = \frac{1}{p} \sum_{k=1}^p \frac{|x_{ik} - x_{jk}|}{x_{ik} + x_{jk}}, (x_{ij} > 0) \quad (10-3.6)$$

在**R**软件中, **dist()**函数给出了各种距离的计算结果, 其调用格式为:

—— **dist()** 的调用格式 ——

```
dist(x, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)
```

说明: **method**表示计算距离的方法, 默认值为**euclidean**(欧氏)距离, **diag**是逻辑变量: 当**diag=TRUE**时, 输出距离矩阵对角线上的距离. **upper**也是逻辑变量: 当**upper=TRUE**时, 输出距离矩阵上三角部分(默认仅输出下三角矩阵)

类与类之间的距离有许多定义方法, 主要有下面七种:

- (1) 类平均法(**average linkage**)
- (2) 重心法(**centroid method**)
- (3) 中间距离法(**median method**)
- (4) 最长距离法(**complete method**)
- (5) 最短距离法(**single method**)
- (6) 离差平方和法(**ward method**)
- (7) **Mcquitty**相似法(**Mcquitty method**)

各类方法计算方式不同, 有学者推荐采用离差平方和法或最短距离法.

10.3.2 R通用程序

利用**R**语言的**hclust()**函数就可完成系统聚类分析, 其基本调用格式如

下:

hclust()的调用格式

```
hclust(d, method = "complete", members=NULL)
```

说明: `d`是由“`dist`”构成的距离结构, `method`是系统聚类的方法(默认地是最长距离法)具体说明见R帮助。

例 10.3.1 设有5个产品, 每个产品测得一项质量指标 x , 其值如下: 1, 2, 4.5, 6, 8, 试用最短距离法、最长距离法、中间距离法、离差平方和法分别对5个产品按质量指标进行分类

解 R程序如下:

```
> x<-c(1, 2, 4.5, 6, 8)
> dim(x)<-c(5, 1)
> d<-dist(x)
> hc1<-hclust(d, "single")
> hc2<-hclust(d, "complete")
> hc3<-hclust(d, "median")
> hc4<-hclust(d, "ward")
> opar<-par(mfrow=c(2, 2))
> plot(hc1, hang=-1);plot(hc2, hang=-1)
> plot(hc3, hang=-1);plot(hc4, hang=-1)
> par(opar)
```

R程序结果见图10.1. 可见, 四种分类方法结果一致, 都将第1, 2个分在一类, 其余在第二类.



例 10.3.2 对例10.2.1中的鸢尾花(Iris)数据进行聚类分析.

解 判别分析中, 我们已知鸢尾花的品种并应用了这些数据. 现在假设我们只知道数据内有三种品种的鸢尾花而不知道每朵花的真正分类, 只能凭借花萼及花瓣的长度和宽度去分成三类, 这就是聚类分析.

R程序如下:

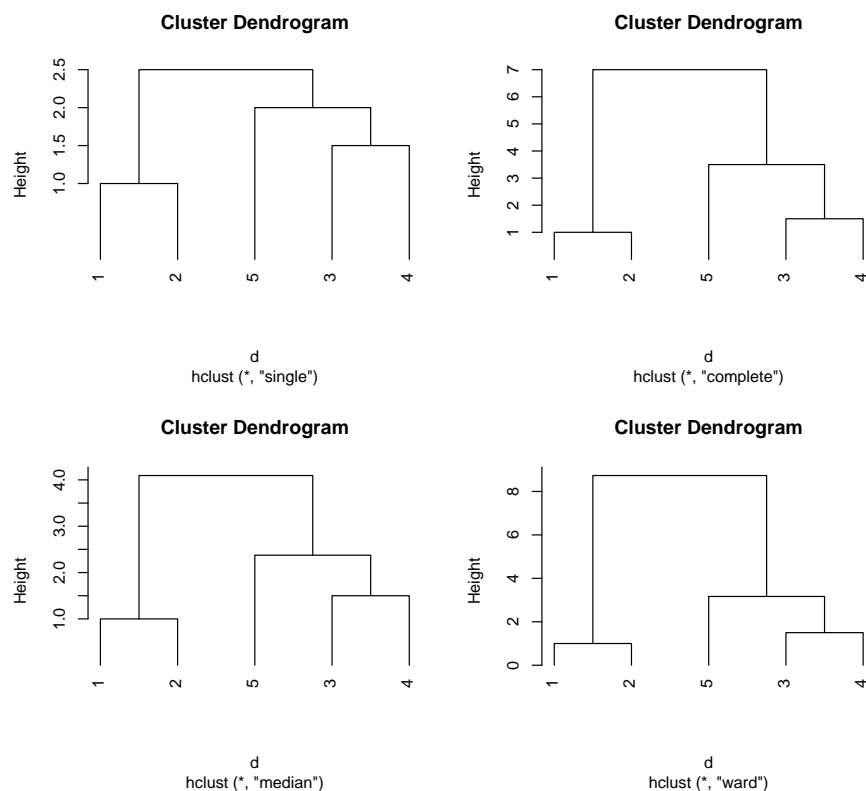


图 10.1 聚类图

```
> data(iris); attach(iris)
> iris.hc<-hclust(dist(iris[,1:4]))
> # plot(iris.hc1, hang = -1)
> plclust(iris.hc1,labels = FALSE, hang=-1)
> re<-rect.hclust(iris.hc1,k=3)
> iris.id <- cutree(iris.hc1,3)
> table(iris.id,Species)
```

程序中我们调用R内置数据iris, 用函数hclust()进行聚类分析, 输出结果保存在iris.hc中, 用函数rect.hclust()按给定的类的个数(或阈值)进行聚类, 并用函数plclust()代替plot()绘制聚类的谱系图(两者使用方法基本相同), 各类用边框界定, 选项labels=FALSE只是为了省去数据的标签. 函数cuttree()将iris.hc输出编制成若干组.

R程序的结果见图10.2和相应的输出.

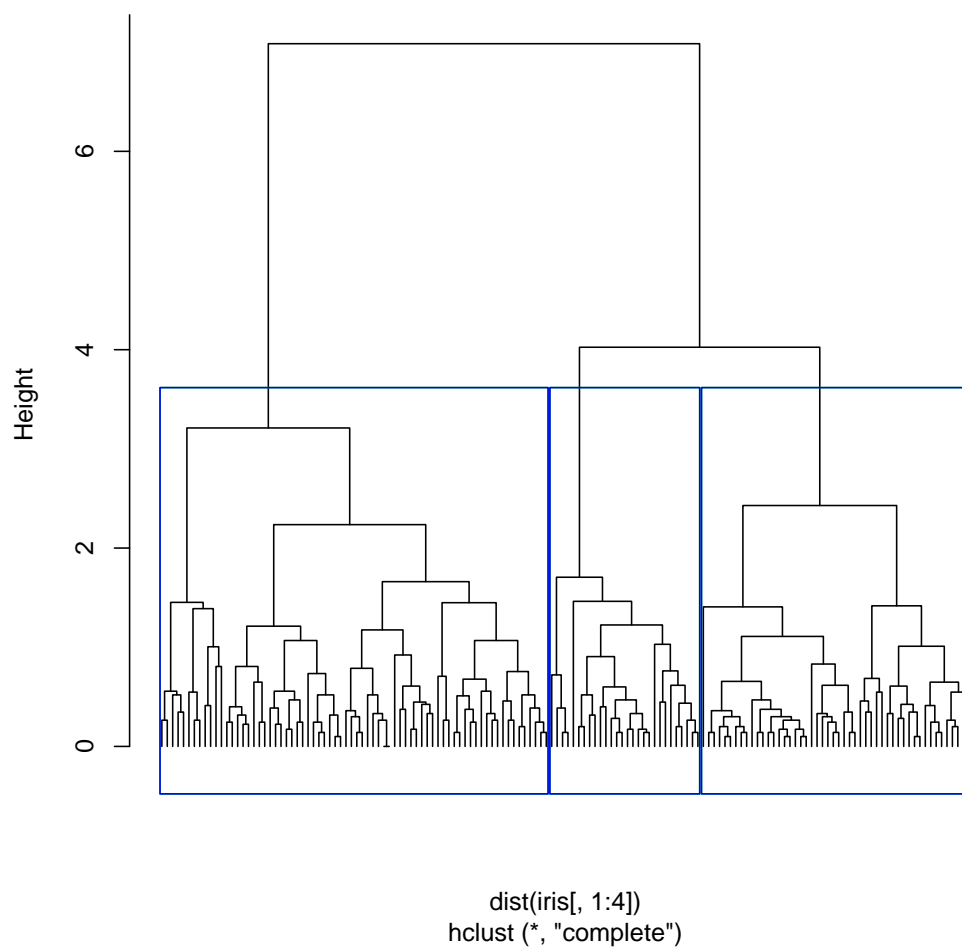


图 10.2 聚类图

Species			
iris.id	setosa	versicolor	virginica
1	50	0	0
2	0	23	49
3	0	27	1

说明: 图10.2为典型的聚类树枝型分类图(Cluster Dendrogram), 它是将两相近(距离最短)的数据向量连接在一起, 然后进一步组合, 直至所有数据都连接在一起; 函数cuttree()将数据iris分类结果iris.hc编为三组, 分别以1, 2, 3表示, 保存在iris.id中. 将iris.id与iris中Species作比较发现, 1应该是setosa类, 2应该是virginica类(因为virginica的个数明显多于versicolor), 3是versicolor. 从聚类的结果来看, 明显与原始数据有着比较大的差异. ■

§10.4 典型相关分析

10.4.1 基本思想

在一元统计分析中, 研究两个随机变量之间的线性相关关系, 可用相关系数(称为简单相关系数); 研究一个随机变量与多个随机变量之间的线性相关关系, 可用复相关系数(称为全相关系数). 1936年Hotelling首先将它推广到研究多个随机变量与多个随机变量之间的相关关系的讨论中, 提出了典型相关分析.

实际问题中, 两组变量之间具有相关关系的问题很多, 例如几种主要产品如猪肉、牛肉、鸡蛋的价格(作为第一组变量) 和相应这些产品的销售量(作为第二组变量)有相关关系; 投资性变量(如劳动者人数、货物周转量、生产建设投资等)与国民收入变量(如工农业国民收入、运输业国民收入、建筑业国民收入等)具有相关关系; 患某种疾病的病人的各种症状程度(第一组变量)和用物理化学方法检验的结果(第二组变量)具有相关关系; 运动员的体力测试指标(如反复横向跳、纵跳、背力、握力等)与运动能力测试指标(如耐力跑、跳远、投球等)之间具有相关关系等等.

典型相关分析就是研究两组变量之间相关关系的一种多元统计方法, 设两组变量用 X_1, X_2, \dots, X_{p_1} 及 $X_{p_1+1}, X_{p_1+2}, \dots, X_{p_1+p_2}$ 表示, 要研究两组变量的相关关系, 一种方法是分别研究 X_i 与 X_j ($i = 1, \dots, p_1$ $j = p_1 + 1, \dots, p_1 + p_2$) 之间的相关关系, 然后列出相关系数表进行分析, 当两组变量较多时, 这样做不仅烦琐, 也不易抓住问题的实际; 另一种方法采用类似主成分分析的做法, 在每一组变量中都选择若干个有代表性的综合指标(变量的线性组合), 通过研究两组的综合指标之间的关系来反映两组变量之间的相关关系. 比如猪肉价格和牛肉价格用 X_1, X_2 表示, 它们的销售量用 X_3, X_4 表示, 研

究它们之间的相关关系, 从经济学观点就是希望构造一个 X_1, X_2 的线性函数 $y = a_{11}X_1 + a_{12}X_2$ 称为价格指数及 X_3, X_4 的线性函数 $y = a_{21}X_3 + a_{22}X_4$ 称为销售指数, 要求它们之间具有最大相关性, 这就是一个典型相关分析问题.

典型相关分析基本思想: 首先在每组变量中找出变量的线性组合, 使其具有最大相关性, 然后再在每组变量中找出第二对线性组合, 使其分别与第一对线性组合不相关, 而第二对本身具有最大的相关性, 如此继续下去, 直到两组变量之间的相关性被提取完毕为止. 有了这样线性组合的最大相关, 则讨论两组变量之间的相关, 就转化为只研究这些线性组合的最大相关, 从而减少研究变量的个数.

典型相关分析是对两组变量(指标)的每一组作为整体考虑的. 因此, 它能够广泛应用于变量群之间的相关分析研究.

设有两组随机变量 $X^{(1)} = (X_1, X_2, \dots, X_{p_1})'$, $X^{(2)} = (X_{p_1+1}, X_{p_1+2}, \dots, X_{p_1+p_2})'$, 记 $p = p_1 + p_2$ 不妨设 $p_1 \leq p_2$, 假定 $X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$ 的协方差阵 $\Sigma > 0$, 均值向量 $\mu = 0$ (否则只要以 $X - \mu$ 代替 X 即可), 相应的将 Σ 剖分为

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

其中 Σ_{11} 是第一组变量的协方差阵, Σ_{12} 是第一组变量与第二组变量的协方差阵, Σ_{22} 是第二组变量的协方差阵. 要研究 $X^{(1)}, X^{(2)}$ 两组变量之间的相关关系, 前面已介绍作两组变量的线性组合, 即

$$U = l_1X_1 + l_2X_2 + \dots + l_{p_1}X_{p_1} \equiv l'X^{(1)},$$

$$V = m_1X_{p_1+1} + m_2X_{p_1+2} + \dots + m_{p_2}X_{p_1+p_2} \equiv m'X^{(2)},$$

其中 $l' = (l_1, l_2, \dots, l_{p_1})$, $m' = (m_1, m_2, \dots, m_{p_2})$ 为任意非零常数向量, 易见:

$$\text{Var}(U) = \text{Var}(l'X^{(1)}) = l'\Sigma_{11}l,$$

$$\text{Var}(V) = \text{Var}(m'X^{(2)}) = m'\Sigma_{22}m,$$

$$\text{Cov}(U, V) = l'\text{Cov}(X_1, X_2)m = l'\Sigma_{12}m,$$

$$\rho_{UV} = \frac{l'\Sigma_{12}m}{\sqrt{l'\Sigma_{11}l}\sqrt{m'\Sigma_{22}m}}.$$

我们寻求 l 与 m 使得 ρ_{UV} 达到最大,但由于随机变量乘以常数时不改变他们的相关系数,为防止不必要的结果重复出现,最好的限制是令 $\text{Var}(U) = l' \Sigma_{11} l = 1$, $\text{Var}(V) = m' \Sigma_{22} m = 1$. 于是我们的问题就成为在约束条件: $\text{Var}(U) = 1$, $\text{Var}(V) = 1$ 下, 寻求 l 与 m 使得 ρ_{UV} 达到最大.

所以典型相关分析研究的是如何选取典型变量的最优组合. 选取的原则是: 在所有的线性组合 U, V 中, 选取典型相关系数最大的 U, V , 即选取 $l^{(1)'}$, $m^{(1)'}$ 使得 $U_1 = l^{(1)' } X^{(1)}$, $V_1 = m^{(1)' } X^{(2)}$ 之间的相关系数达到最大(在所有 U, V 中), 然后选取 $l^{(2)'}$, $m^{(2)'}$ 使得 $U_2 = l^{(2)' } X^{(1)}$, $V_2 = m^{(2)' } X^{(2)}$ 之间的相关系数在与 U_1, V_1 不相关的组合 U, V 中达到最大(第二高的相关). 如此继续下去, 直到选取出所有分别与 U_1, U_2, \dots, U_{k-1} 和 V_1, V_2, \dots, V_{k-1} 都不相关的线性组合 U_k, V_k 为止, 此时 k 为两组原始变量中个数较少的那个数. 典型变量 U_1 和 V_1, U_2 和 V_2, \dots, U_k 和 V_k 是根据它们的相关系数由大到小逐对提取的, 直到两组变量之间的相关性被分解完毕为止.

10.4.2 R通用程序

利用R语言的`cancor()`函数就可完成典型相关分析. 其基本调用格式如下:

cancor() 的调用格式

```
cancor(x, y, xcenter = TRUE, ycenter = TRUE)
```

说明: x, y 是两组变量的数据矩阵, $xcenter$ 和 $ycenter$ 是逻辑变量, TRUE表示将数据中心化(默认选项), 具体说明见R帮助.

例 10.4.1 研究投资性变量与反映国民经济变量之间的相关关系. 投资性变量选6个, 分别为 X_1, X_2, \dots, X_6 , 反映国民经济的变量选5个, 分别为 Y_1, Y_2, \dots, Y_5 . 抽取从1975—2002年共计28年的统计数据, 如表10.4, 采用典型相关分析的方法来分析投资性变量与反映国民经济的变量的相关性.

解 R程序如下:

```
> invest=read.table("D:/Rdata/invest.txt")
> names(invest)=c("x1", "x2", "x3", "x4", "x5", "x6",
                  "y1", "y2", "y3", "y4", "y5")
> ca<-cancor(invest[, 1:6], invest[, 7:11])
```

R程序结果:

表 10.4 1975—2002年的投资性变量与反映国民经济的变量

序列	X_1	X_2	X_3	X_4	X_5	X_6	Y_1	Y_2	Y_3	Y_4	Y_5
1	173.28	93.62	60.10	86.72	38.97	27.51	75.3	117.4	74.6	61.8	4508
2	172.09	92.83	60.38	87.39	38.62	27.82	76.7	120.1	77.1	66.2	4469
3	171.46	92.73	59.74	85.59	38.83	27.46	75.8	121.8	75.2	65.4	4398
4	170.08	92.25	58.04	85.92	38.33	27.29	76.1	115.1	73.8	61.3	4068
5	170.61	92.36	59.67	87.46	38.38	27.14	72.9	119.4	77.5	67.1	4339
6	171.69	92.85	59.44	87.45	38.19	27.10	72.7	116.2	74.6	59.3	4393
7	171.46	92.93	58.70	87.06	38.58	27.36	76.5	117.9	75.0	68.3	4389
8	171.60	93.28	59.75	88.03	38.68	27.22	75.2	115.1	74.1	63.2	4306
9	171.60	92.26	60.50	87.63	38.79	6.63	74.7	117.4	78.3	68.3	4395
10	171.16	92.62	58.72	87.11	38.19	27.18	73.2	113.2	72.5	51.0	4462
11	170.04	92.17	56.95	88.08	38.24	27.65	77.8	116.9	76.9	65.6	4181
12	170.27	91.94	56.00	84.52	37.16	26.81	76.4	113.6	74.3	65.6	4232
13	170.61	92.50	57.34	85.61	38.52	27.36	76.4	116.7	74.3	61.2	4305
14	171.39	92.44	58.92	85.37	38.83	26.47	74.9	113.1	74.0	61.2	4276
15	171.83	92.79	56.85	85.35	38.58	27.03	78.7	112.4	72.9	61.4	4067
16	171.36	92.53	58.39	87.09	38.23	27.04	73.9	118.4	73.0	62.3	4421
17	171.24	92.61	57.69	83.98	39.04	27.07	75.7	116.3	74.2	51.8	4284
18	170.49	92.03	57.56	87.18	38.54	27.57	72.5	114.8	71.0	55.1	4289
19	169.43	91.67	57.22	83.87	38.41	26.60	76.7	117.5	72.7	51.6	4097
20	168.57	91.40	55.96	83.02	38.74	26.97	77.0	117.9	71.6	52.4	4063
21	170.43	92.38	57.87	84.87	38.78	27.37	76.0	116.8	72.3	58.0	4334
22	169.88	91.89	56.87	86.34	38.37	27.19	74.2	115.4	73.1	60.4	4301
23	167.94	90.91	55.97	86.77	38.17	27.16	76.2	110.9	68.5	56.8	4141
24	168.82	91.30	56.07	85.87	37.61	26.67	77.2	113.8	71.0	57.5	3905
25	168.02	91.26	55.28	85.63	39.66	28.07	74.5	117.2	74.0	63.8	3943
26	167.87	90.96	55.79	84.92	38.20	26.53	74.3	112.3	69.3	50.2	4195
27	168.15	91.50	54.56	84.81	38.44	27.38	77.5	117.4	75.3	63.6	4039
28	168.99	91.52	55.11	86.23	38.30	27.14	77.7	113.3	72.1	52.8	4238

\$cor

```
[1] 0.8743062 0.7373122 0.5104993 0.3541742 0.1510162
```

```
$xcoef
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
x1	0.07907994	-0.148819922	0.10698911	-0.023654480	0.13247040	0.510516673
x2	-0.06231142	-0.005330072	0.16195009	0.002344311	-0.47379144	-0.965367730
x3	0.05901271	0.181098314	-0.08095237	0.045803287	0.12588000	-0.078796849
x4	0.02720459	-0.142360745	-0.11291982	-0.015211819	-0.09859752	0.003270612
x5	-0.05046629	-0.026272162	-0.10552122	0.250540745	-0.43012143	0.205369791
x6	0.09818550	0.408743145	0.07539201	-0.489439843	0.25175511	0.209683441

```
$ycoef
```

	[,1]	[,2]	[,3]	[,4]	[,5]
y1	-0.0097969228	-0.0157649468	0.1263997157	-4.494590e-02	0.0042832480
y2	-0.0064201995	0.0954303161	0.0052805529	-1.885232e-02	0.0249264752
y3	0.0179056849	-0.0279923087	0.0194826255	9.233173e-02	-0.1181715021
y4	0.0101074931	-0.0082248457	-0.0202537565	-4.602875e-02	0.0097856402
y5	0.0009083347	-0.0003598887	0.0007780556	-8.745514e-06	0.0008253158

```
$xcenter
```

	x1	x2	x3	x4	x5	x6
	170.37000	92.19750	57.69429	86.06679	38.47786	27.17107

```
$ycenter
```

	y1	y2	y3	y4	y5
	75.59643	116.01071	73.68571	60.11429	4251.35714

结果说明:

- 1) \$cor给出了典型相关系数; \$xcoef是对应于数据X的系数, 即为关于数据X的典型载荷; \$ycoef为关于数据Y的典型载荷; \$xcenter与\$ycenter是数据X与Y的中心, 即样本均值;
- 2) 对于该问题, 第一对典型变量的表达式为

$$\begin{aligned} U_1 &= 0.079X_1 - 0.062X_2 + 0.059X_3 + 0.027X_4 - 0.050X_5 + 0.098X_6 \\ V_1 &= -0.010Y_1 - 0.006Y_2 + 0.0179Y_3 + 0.010Y_4 + 0.001Y_5 \end{aligned}$$

第一对典型变量的相关系数为0.8743062.

可以进行典型相关系数的显著性检验, 经检验也只有第一组典型变量.



§10.5 对应分析

对应分析(Correspondence Analysis)又称为相应分析, 是1970年由法国统计学家J.P.Beozecri提出来的. 对应分析是因子分析的进一步推广, 该方法已成为多元统计分析中同时对样品和变量进行分析, 从而研究多变量内部关系的重要方法, 它是在R型和Q型因子分析基础上发展起来的一种多元统计方法. 而且我们研究样本之间或指标之间的关系, 归根结底是为了研究样本与指标之间的关系, 而因子分析没有办法做到这一点, 对应分析则是为解决这个问题而出现的统计分析方法.

10.5.1 基本思想

由于R型因子分析和Q型因子分析都是反应一个整体的不同侧面, 因而它们之间一定存在着内在的联系. 对应分析就是通过对应变换后的过渡矩阵 Z 将两者有机地结合起来.

假设有 n 个样本, 每个样本有 p 个指标, 原始数据矩阵用 $X_{n \times p}$ 来表示. 研究指标或样本之间的关系是分别通过研究它们的协方差矩阵 $A_{p \times p}$ 或相似矩阵 $B_{n \times n}$ 进行的, 实际上用到的只是这些矩阵的特征根和特征向量, 因此, 如能由矩阵 $A_{p \times p}$ 的特征根和特征向量直接得出矩阵 $B_{n \times n}$ 的特征根和特征向量, 而不必计算相似矩阵 $B_{n \times n}$, 则就解决了当样本数很大时做Q型因子分析计算上的困难.

对应分析就是利用降维的思想, 通过一个过渡矩阵 Z 将上述二者有机地结合起来, 具体地说, 首先给出变量点的协差阵 $A = Z'Z$ 和样品点的协差阵 $B = ZZ'$. 由于 $A = Z'Z$ 和 $B = ZZ'$ 有相同的非零特征根记为 $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_m, 0 \leq m \leq \min(n, p)$. 如果 A 的特征根 λ_i 对应的特征向量为 U_i , 则 B 的特征根 λ_i 对应的特征向量就是 $ZU_i = V_i$, 根据这个结论(证明省略) 就可以很方便的借助R型因子分析而得到Q型因子分析的结果. 因此求出 A 的特征根和特征向量后就很容易地写出变量协差阵对应的因子载荷阵, 记为 F , 则

$$F = \begin{bmatrix} u_{11}\sqrt{\lambda_1} & u_{12}\sqrt{\lambda_1} & \cdots & u_{1m}\sqrt{\lambda_m} \\ u_{21}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_1} & \cdots & u_{2m}\sqrt{\lambda_m} \\ \cdots & \cdots & \cdots & \cdots \\ u_{p1}\sqrt{\lambda_1} & u_{p2}\sqrt{\lambda_1} & \cdots & u_{pm}\sqrt{\lambda_m} \end{bmatrix}$$

这样一来样品点协差阵B对应的因子载荷阵记为G, 则

$$G = \begin{bmatrix} v_{11}\sqrt{\lambda_1} & v_{12}\sqrt{\lambda_1} & \cdots & v_{1m}\sqrt{\lambda_m} \\ v_{21}\sqrt{\lambda_1} & v_{22}\sqrt{\lambda_1} & \cdots & v_{2m}\sqrt{\lambda_m} \\ \cdots & \cdots & \cdots & \cdots \\ v_{n1}\sqrt{\lambda_1} & v_{n2}\sqrt{\lambda_1} & \cdots & v_{nm}\sqrt{\lambda_m} \end{bmatrix}$$

由于A和B具有相同的非零特征根, 而这些特征根又正是各个公共因子的方差, 因此可以用相同的因子轴同时表示变量点和样品点, 即把变量点和样品点同时反映在具有相同坐标轴的因子平面上, 以便对变量点和样品点一起考虑进行分类. 那么矩阵 $A_{p \times p}$ 与矩阵 $B_{n \times n}$ 是否存在必然的联系呢? 这种联系的确是存在的, 因为 $A_{p \times p}$ 和 $B_n \times n$ 都来自于同样的原始数据 $X_{n \times p}$, $X_{n \times p}$ 中的每一个元素 x_{ij} 都具有双重含义, 同时代表指标和样本. 实际上指标与样本是不可分割的, 指标的特征如均值、协方差等是通过指标在不同样本上的取值来表现的, 而样本的特征如样本属于哪一类型, 正是通过其在不同指标上的取值来表现的. 但是, 要由矩阵 $A_{p \times p}$ 的特征根和特征向量直接求出矩阵 $B_{n \times n}$ 的特征根和特征向量还是有困难的, 因为 $A_{p \times p}$ 与 $B_{n \times n}$ 的阶数不一样, 一般来说, 其非零特征根也不相等. 如果能将原始数据矩阵X进行某种变形后成为Z, 使得 $A = Z'Z$ $B = ZZ'$, 由线性代数可知, $Z'Z$ 和 ZZ' 有相同的非零特征根, 记为 $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_m, 0 \leq m \leq \min(n, p)$, 设 u_1, \cdots, u_γ 为对应于特征根 $\lambda_1, \cdots, \lambda_\gamma$ 的A的特征向量, 则有

$$Au_j = Z'Z u_j = \lambda_j u_j.$$

将上式两边左乘Z, 得

$$ZZ'Z u_j = Z\lambda_j u_j = \lambda_j Z u_j$$

即

$$B(Zu_j) = \lambda_j(Zu_j)$$

上式表明, Zu_j 为对应于特征根 λ_j 的 B 的特征向量. 换句话说, 当 u_j 为对应于 λ_j 的 A 的特征向量时, 则 Zu_j 就是对应于 λ_j 的 B 的特征向量. 这样就建立起了因子分析中 R 型与 Q 型的关系, 而且使计算变得方便多了.

综上所述, 若将原始数据矩阵变换 X 为 Z 时, 则指标和样本的协方差阵可分别表示为 $A = Z'Z$ 和 $B = ZZ'$, A 和 B 具有相同的非零特征根, 相应的特征向量具有很密切的关系, 这样就可很方便地从 R 型因子分析出发而直接得到 Q 型因子分析的结果, 从而克服了大样本时做 Q 型因子分析计算上的困难. 又由于 A 和 B 具有相同的非零特征根, 而这些特征根正是各个因子所提供的方差, 那么在 p 维指标空间 R^p 中和 n 维样本空间 R^n 中各个主因子在总方差中所占的比重就完全相同, 即指标空间中的第一主因子也是样本空间中的第一主因子, 依次类推. 这样就可利用相同的因子轴去同时表示指标和样本, 将指标和样本同时反映在有相同坐标轴的因子轴的因子平面上. 因此, 对应分析的关键在于如何将 X 变换成 Z .

1970年, 法国统计学家 J.P. Beocri 提出了上述求 Z 的方法. 基本步骤为: X 标准化处理 \rightarrow 求指标的均值 (可证明亦是样本的均值) \rightarrow 求协方差矩阵 $A \rightarrow$ 将 A 变形为 $A = Z'Z \rightarrow Z$

10.5.2 R 通用程序

首先我们要用指令 `library(MASS)` 加载 `MASS` 宏包, 再用 `corresp()` 函数就可完成简单对应分析, 其基本调用格式如下:

`corresp(x, nf = 1, ...)`

说明: x 是数据矩阵, $nf = 1$ 表示计算因子个数, 具体说明见 **R** 帮助.

例 10.5.1 (妇女就业问题) 利用 90 年代初期对某市若干个郊区已婚妇女的调查资料, 主要调查她们对“应该男人在外工作, 妇女在家操持家务”的态度, 依据文化程度和就业观点两个变量进行分类汇总, 数据如表 10.5.

解 **R** 程序如下:

```
> x.df=data.frame(HighlyFor=c(2, 6, 41, 72, 24),
                  For=c(17, 65, 220, 224, 61),
                  Against=c(17, 79, 327, 503, 300),
                  HighlyAgainst=c(5, 6, 48, 47, 41))
```

表 10.5 妇女就业问题调查

文化程度	就业观点			
	非常同意	同意	不同意	非常不同意
小学以下	2	17	17	5
小学	6	65	79	6
初中	41	220	327	48
高中	72	224	503	47
大学	24	61	300	41

```

> rownames(x.df)<-c("BelowPrimary", "Primary",
                    "Secondary", "HighSchool","College")
> library(MASS)
> biplot(corresp(x.df, nf=2))

```

说明: biplot作出像因子分析的载荷图那样的, 这样可以直观地来展示两个变量各个水平之间的关系.

R程序结果如图10.3.

结果说明:

- 1) 对于该图, 主要看横坐标的两种点(就业观点与文化程度)的距离, 纵坐标的距离对于分析贡献意义不大.
- 2) 对于该图可以看出对该观点持赞同态度的是小学以下, 小学, 初中, 而大学文化程度的妇女主要持不同意或者非常不同意的观点, 高中文化程度的持有非常不赞同或者非常同意两种观点.



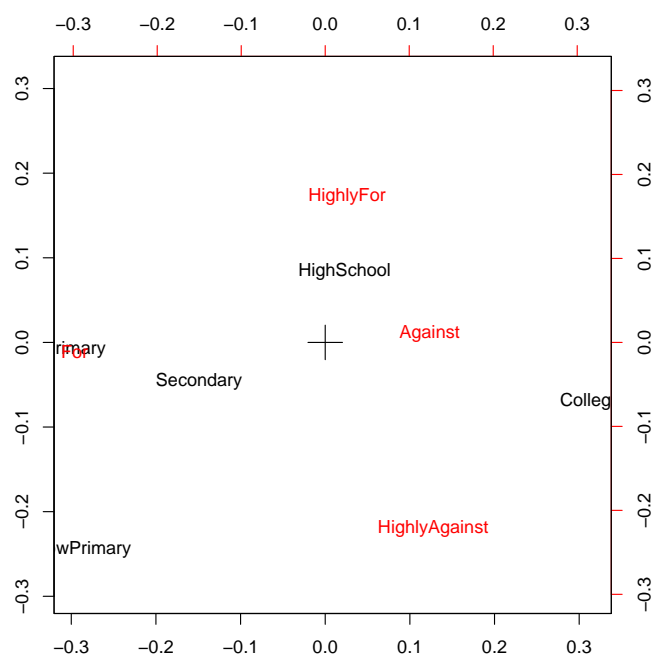


图 10.3 .

第十章习题

10.1 服装定型的分类问题: 为解决服装定型分类问题, 对128个成年人的身材进行了测量, 每人各测得16项指标: 身高(X_1), 坐高(X_2), 胸围(X_3), 头高(X_4), 裤长(X_5), 下档(X_6), 手长(X_7), 领围(X_8), 前胸(X_9), 后背(X_{10}), 肩厚(X_{11}), 肩宽(X_{12}), 袖长(X_{13}), 肋围(X_{14}), 腰围(X_{15}), 腿肚(X_{16}). 16项指标的相关阵见第363页表10.6, 试从相关阵出发进行主成分分析.

10.2 第十章习题犯罪问题的主成分分析: 本例的输入资料文件是美国50个州(state) 在七种犯罪项目上的发生频率. 这七种罪分别是: 谋杀(MURDER), 强暴(RAPE), 抢劫(ROBBERY), 骚扰(ASSAULT), 夜间偷窃(简称夜盗, BURGLARY), 盗窃(LARCENY)及偷车(AUTO), 数据如表10.7. 试图用主成分分析降维处理.

表 10.7: 各州犯罪数据

州名	谋杀	强暴	抢劫	骚扰	夜盗	盗窃	偷车
Alabama	14.2	25.2	96.8	278.3	1135.5	1881.9	280.7
Alaska	10.8	51.6	96.8	284.0	1331.7	3369.8	753.3
Arizona	9.5	34.2	138.2	312.3	2346.1	4467.4	439.5
Arkansas	8.8	27.6	83.2	203.4	972.6	1862.1	183.4
California	11.5	49.4	287.0	358.0	2139.4	3499.8	663.5
Colorado	6.3	42.0	170.7	292.9	1935.2	3903.2	477.1
Connecticut	4.2	16.8	129.5	131.8	1346.0	2620.7	593.2
Delaware	6.0	24.9	157.0	194.2	1682.6	3678.4	467.0
Florida	10.2	39.6	187.9	449.1	1859.9	3840.5	351.4
Georgia	11.7	31.1	140.5	256.5	1351.1	2170.2	297.9
Hawaii	7.2	25.5	128.0	64.1	1911.5	3920.4	489.4
Idaho	5.5	19.4	39.6	172.5	1050.8	2599.6	237.6

续下页

各州犯罪数据 (续表)

州名	谋杀	强暴	抢劫	骚扰	夜盗	盗窃	偷车
Illinois	9.9	21.8	211.3	209.0	1085.0	2828.5	528.6
Indiana	7.4	26.5	123.2	153.5	1086.2	2498.7	377.4
Iowa	2.3	10.6	41.2	89.8	812.5	2685.1	219.9
Kansas	6.6	22.0	100.7	180.5	1270.4	2739.3	244.3
Kentucky	10.1	19.1	81.1	123.3	872.2	1662.1	245.4
Louisiana	15.5	30.9	142.9	335.5	1165.5	2469.9	337.7
Maine	2.4	13.5	38.7	170.0	1253.1	2350.7	246.9
Maryland	8.0	34.8	292.1	358.9	1400.0	3177.7	428.5
Massachusetts	3.1	20.8	169.1	231.6	1532.2	2311.3	1140.1
Michigan	9.3	38.9	261.9	274.6	1522.7	3159.0	545.5
Minnesota	2.7	19.5	85.9	85.8	1134.7	2559.3	343.1
Mississippi	14.3	19.6	65.7	189.1	915.6	1239.9	144.4
Missouri	9.6	28.3	189.0	233.5	1318.3	2424.2	378.4
Montana	5.4	16.7	39.2	156.8	804.9	2773.2	309.2
Nebraska	3.9	18.1	64.7	112.7	760.0	2316.1	249.1
Nevada	15.8	49.1	323.1	355.0	2453.1	4212.6	559.2
New Hampshire	3.2	10.7	23.2	76.0	1041.7	2343.9	293.4
New Jersey	5.6	21.0	180.4	185.1	1435.8	2774.5	511.5
New Mexico	8.8	39.1	109.6	343.4	1418.7	3008.6	259.5
New York	10.7	29.4	472.6	319.1	1728.0	2782.0	745.8

续下页

各州犯罪数据 (续表)

州名	谋杀	强暴	抢劫	骚扰	夜盗	盗窃	偷车
North Carolina	10.6	17.0	61.3	318.3	1154.1	2037.8	192.1
North Dakota	0.9	9.0	13.3	43.8	446.1	1843.0	144.7
Ohio	7.8	27.3	190.5	181.1	1216.0	2696.8	400.4
Oklahoma	8.6	29.2	73.8	205.0	1288.2	2228.1	326.8
Oregon	4.9	39.9	124.1	286.9	1636.4	3506.1	388.9
Pennsylvania	5.6	19.0	130.3	128.0	877.5	1624.1	333.2
Rhode Island	3.6	10.5	86.5	201.0	1489.5	2844.1	791.4
South Carolina	11.9	33.0	105.9	485.3	1613.6	2342.4	245.1
South Dakota	2.0	13.5	17.9	155.7	570.5	1704.4	147.5
Tennessee	10.1	29.7	145.8	203.9	1259.7	1776.5	314.0
Texas	13.3	33.8	152.4	208.2	1603.1	2988.7	397.6
Utah	3.5	20.3	68.8	147.3	1171.6	3004.6	334.5
Vermont	1.4	15.9	30.8	101.2	1348.2	2201.0	265.2
Virginia	9.0	23.3	92.1	165.7	986.2	2521.2	226.7
Washington	4.3	39.6	106.2	224.8	1605.6	3386.9	360.3
West Virginia	6.0	13.2	42.2	90.9	597.4	1341.7	163.3
Wisconsin	2.8	12.9	52.2	63.7	846.9	2614.2	220.7
Wyoming	5.4	21.9	39.7	173.9	811.6	2772.2	282.0

10.3 考试成绩分析: 某年级44名学生的期末考试共有5门课程, 有的用闭卷, 有的用开卷, 数据如表10.8. 试用因子分析方法分析这组数据.

表 10.8: 考试成绩分析数据

力学(闭)	物理(闭)	代数(开)	分析(开)	统计(开)
X_1	X_2	X_3	X_4	X_5
77	82	67	67	81
75	73	71	66	81
63	63	65	70	63
51	67	65	65	68
62	60	58	62	70
52	64	60	63	54
50	50	64	55	63
31	55	60	57	73
44	69	53	53	53
62	46	61	57	45
44	61	52	62	46
12	58	61	63	67
54	49	56	47	53
44	56	55	61	36
46	52	65	50	35
30	69	50	52	45
40	27	54	61	61
36	59	51	45	51
46	56	57	49	32
42	60	54	49	33
23	55	59	53	44
41	63	49	46	34

续下页

考试成绩分析数据 (续表)

力学(闭)	物理(闭)	代数(开)	分析(开)	统计(开)
X_1	X_2	X_3	X_4	X_5
63	78	80	70	81
55	72	63	70	68
53	61	72	64	73
59	70	68	62	56
64	72	60	62	45
55	67	59	62	44
65	63	58	56	37
60	64	56	54	40
42	69	61	55	45
31	49	62	63	62
49	41	61	49	64
49	53	49	62	47
54	53	46	59	44
18	44	50	57	81
32	45	49	57	64
46	49	53	59	37
31	42	48	54	68
56	40	56	54	35
45	42	55	56	40
40	63	53	54	25
48	48	49	51	37
46	52	53	41	40

10.4 医药行业数据分析: 数据集(见表10.9)中的数据是全国医药行业20个

企业1980-1982年三年平均效益的几个数据, 总产值/消耗(X_1), 净产值/工资(X_2), 盈利/资金占用(X_3), 销售收入/成本(X_4), 试用因子分析方法找出这4个变量的公因子, 并进行合理的解释.

表 10.9: 医药行业效数据

总产值/消耗 X_1	净产值/工资 X_2	盈利/资金占用 X_3	销售收入/成本 X_4
1.611	10.59	0.69	1.67
1.429	9.44	0.61	1.50
1.447	5.97	0.24	1.25
1.572	10.72	0.75	1.71
1.483	10.99	0.75	1.44
1.371	6.46	0.41	1.31
1.665	10.51	0.53	1.52
1.403	6.11	0.17	1.32
2.620	21.51	1.40	2.59
2.033	24.15	1.80	1.89
2.015	26.86	1.93	2.02
1.501	9.74	0.87	1.48
1.578	14.52	1.12	1.47
1.735	14.64	1.21	1.91
1.453	12.88	0.87	1.52
1.765	17.94	0.89	1.40
1.532	29.42	2.52	1.80
1.488	9.23	0.81	1.45
2.586	16.07	0.82	1.83
1.992	21.63	1.01	1.89

10.5 胃癌的鉴别: 表10.10是从病例中随即抽取的部分资料. 这里有3个类别(group): 胃癌(ca)、萎缩性胃炎(ga)和非胃炎患者(non). 从每个总体抽5个病人, 每人化验4项生化指标: 血清铜蛋白(X_1)、蓝色反应(X_2)、尿乙酸(X_3)和中性硫化物(X_4). 试对胃癌检验的生化指标值用Fisher 判别的方法进行判别归类.

表 10.10: 胃癌检验的生化指标值

类别	序号	血清铜蛋白 X_1	蓝色反应 X_2	尿乙酸 X_3	中性硫化物 X_4
胃癌患者	1	228	134	20	11
	2	245	134	10	40
	3	200	167	12	27
	4	170	150	7	8
	5	100	167	20	14
萎缩性胃炎患者	6	225	125	7	14
	7	130	100	6	12
	8	150	117	7	6
	9	120	133	10	26
	10	160	100	5	10
非胃炎患者	11	185	115	5	19
	12	170	125	6	4
	13	165	142	5	3
	14	135	108	2	12
	15	100	117	7	2

10.6 设有6个产品, 每个产品测得一项质量指标 X , 其值如下: 1, 2, 4, 6, 9, 11. 试对6个产品按质量指标进行分类, 试用各种系统聚类方法进行分析, 然后比较之.

10.7 生活消费水平聚类分析: 表10.11中的资料是我国16个地区农民1982年

支出情况的抽样调查的汇总资料, 每个地区都调查了反映每人平均生活消费支出情况的六个指标, 分别是食品(X_1), 衣着(X_2), 燃料(X_3), 住房(X_4), 生活用品及其他(X_5), 文化生活服务支出(X_6). 试利用调查资料对16个地区进行分类.

表 10.11: 中国农民1982年各类支出

地区 (area)	食品 X_1	衣着 X_2	燃料 X_3	住房 X_4	生活用品及 其他 X_5	文化生活服务 支出 X_6
北京	190.33	43.77	9.73	60.54	49.01	9.04
天津	135.20	36.40	10.47	44.16	36.49	3.94
河北	95.21	22.83	9.30	22.44	22.81	2.80
山西	104.78	25.11	6.40	9.89	18.17	3.25
内蒙	128.41	27.63	8.94	12.58	23.99	3.27
辽宁	145.68	32.83	17.79	27.29	39.09	3.47
吉林	159.37	33.38	18.37	11.81	25.29	5.22
黑龙江	116.22	29.57	13.24	13.76	21.75	6.04
上海	221.11	38.64	12.53	115.65	50.82	5.89
江苏	144.98	29.12	11.67	42.60	27.30	5.74
浙江	169.92	32.75	12.72	47.12	34.35	5.00
安徽	153.11	23.09	15.62	23.54	18.18	6.39
福建	144.92	21.26	16.96	19.52	21.75	6.73
江西	140.54	21.50	17.64	19.19	15.97	4.94
山东	115.84	30.26	12.20	33.61	33.77	3.85
河南	101.18	23.26	8.46	20.20	20.50	4.30

10.8 矿产数据的典型相关分析: 为了了解某矿区下部矿Pt(铂), Pd(钯)与Cu(铜), Ni(镍)的共生组合规律, 我们从其钻孔中取出27个样品(数据见表10.12). 试用典型相关分析研究Pt(铂), Pd(钯)与Cu(铜), Ni(镍)的相关关系.

表 10.12: 矿区下部的矿产数据

序号	Pt(铂)	Pd(钯)	Cu(铜)	Ni(镍)
	X_1	X_2	X_3	X_4
1	0.14	0.30	0.03	0.14
2	0.20	0.50	0.14	0.22
3	0.06	0.11	0.03	0.02
4	0.07	0.11	0.04	0.13
5	0.12	0.22	0.06	0.12
6	0.52	0.87	0.19	0.20
7	0.23	0.47	0.14	0.10
8	1.19	0.38	0.09	0.11
9	0.37	0.66	0.14	0.15
10	0.36	0.60	0.14	0.15
11	0.42	0.77	0.17	0.10
12	0.35	0.85	0.30	0.19
13	0.50	0.87	0.23	0.22
14	0.56	1.15	0.29	0.28
15	0.43	0.90	0.13	0.22
16	0.47	0.97	0.26	0.22
17	0.49	0.79	0.21	0.20
18	0.47	0.77	0.51	0.22
19	0.40	0.88	0.33	0.19
20	0.66	1.30	0.21	0.30
21	0.63	1.30	0.45	0.28
22	0.52	1.43	0.31	0.23

续下页

矿区下部的矿产数据 (续表)

序号	Pt(铂) X_1	Pd(钯) X_2	Cu(铜) X_3	Ni(镍) X_4
23	0.44	0.87	0.17	0.25
24	0.03	0.07	0.05	0.08
25	0.20	0.28	0.04	0.08
26	0.04	0.10	0.11	0.07
27	0.17	0.28	0.15	0.09

10.9 遗传数据的典型相关分析: 表10.13列举了25个家庭的成年长子和次子的头长和头宽, 可以想象, 长子和次子之间有相当的相关性. 试对长子和次子之间作出典型相关分析.

表 10.13: 长子和次子的遗传数据

长子头长 X_1	长子头宽 X_2	次子头长 Y_1	次子头宽 Y_2
191	155	179	145
195	149	201	152
181	148	185	149
183	153	188	149
176	144	171	142
208	157	192	152
189	150	190	149
197	159	189	152
188	152	197	159
192	150	187	151
179	158	186	148

续下页

长子和次子的遗传数据(续表)

长子头长	长子头宽	次子头长	次子头宽
X_1	X_2	Y_1	Y_2
183	147	174	147
174	150	185	152
190	159	195	157
188	151	187	158
163	137	161	130
195	155	183	158
186	153	173	148
181	145	182	146
175	140	165	137
192	154	185	152
174	143	178	147
176	139	176	143
197	167	200	158
190	163	187	150

10.10 农业生产的典型相关分析: 对表10.14中给出的2001年全国30个省市自治区农业产量(主要是粮食、油料)与农业投入(农作物总播种面积、有效灌溉面积、化肥施用量、农业机械总动力)作典型相关分析.

表 10.14: 2001年全国30个省市自治区农业产量

地区	粮食产量 (万吨)	油料产量 (万吨)	农作物总 播种面积	有效灌 溉面积	化肥施用量 (万吨)	农业机械 总动力
北京	104.9	4.3	386.4	322.7	15.7	395.0
天津	143.3	3.9	544.5	354.3	17.3	603.3

续下页

2001年全国30个省市自治区农业产量(续表)

地区	粮食产量 (万吨)	油料产量 (万吨)	农作物总 播种面积	有效灌 溉面积	化肥施用量 (万吨)	农业机械 总动力
河北	2491.8	153.8	8990.8	4485.4	273.4	7244.4
山西	692.1	18.1	3672.3	1104.3	84.9	1767.5
内蒙古	1239.1	80.6	5707.3	2472.3	79.3	1423.6
辽宁	1394.4	46.3	3964.8	1482.8	109.8	1401.3
吉林	1953.4	34.3	4890.1	1382.6	114.1	1096.5
黑龙江	2651.7	36.3	9989.2	2090.4	123.2	1648.3
上海	151.4	12.8	490.9	280.6	20.3	133.9
江苏	2942.1	232.5	7777.4	3900.0	338.0	2957.9
浙江	1072.7	58.2	3245.9	1400.3	90.3	2017.2
安徽	2500.3	298.8	8733.1	3228.7	280.7	3165.0
福建	817.3	26.1	2713.1	942.4	117.4	888.8
江西	1600.0	90.5	5534.7	1897.5	109.7	1002.0
山东	3720.6	377.3	11266.1	4836.1	428.6	7689.6
河南	4119.9	362.6	13127.7	4766.0	441.7	6078.7
湖北	2138.5	279.4	7489.0	2027.9	245.3	1469.2
湖南	2700.3	137.4	7931.7	2676.3	184.3	2358.0
广东	1600.1	80.9	5193.1	1447.1	195.1	1760.
广西	1511.4	57.2	6288.1	1519.6	168.1	1552.4
海南	195.8	10.3	871.7	180.8	27.0	212.2
重庆	1023.5	30.0	3555.9	631.9	72.6	628.1
四川	2926.5	181.0	9571.5	2533.0	212.0	1735.1
贵州	1100.3	71.3	4650.7	659.8	70.0	647.9
云南	1486.3	27.7	5929.6	1424.3	120.0	1397.8

续下页

2001年全国30个省市自治区农业产量(续表)

地区	粮食产量 (万吨)	油料产量 (万吨)	农作物总 播种面积	有效灌 溉面积	化肥施用量 (万吨)	农业机械 总动力
西藏	98.3	4.4	230.9	154.4	3.0	123.2
陕西	976.6	37.5	4331.9	1314.1	131.1	1099.8
甘肃	753.2	38.4	3688.9	982.3	66.1	1122.0
青海	103.2	23.0	529.0	208.3	7.2	264.7
宁夏	274.8	7.3	1007.6	405.4	24.6	407.6
新疆	780.0	42.6	3404.1	3138.1	83.3	880.9

10.11 城镇居民消费支出结构对应分析: 选取8个反映城镇居民消费支出结构的指标: X_1 —食品支出比重; X_2 —衣着支出比重; X_3 —家庭设备用品及服务支出比重; X_4 —医疗保健支出比重; X_5 —交通和通讯支出比重; X_6 —娱乐教育文化服务支出比重; X_7 —居住支出比重; X_8 —杂项商品支出比重. 根据《2000年统计年鉴》的资料(见表10.15), 进行对应分析.

表 10.15: 城镇居民消费支出结构

地区	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
京	39.5	9.7	10	6.8	6.2	15.2	6.4	6.1
津	42	8.5	11.9	5.2	4.9	12.6	9.8	5.2
冀	37.1	12.8	9	7.1	6.8	13.4	9.1	4.7
晋	40.3	13.7	8.3	6	5.8	11.9	8.1	6.1
内	37.6	15.1	7.3	5.5	7.2	13.3	8.3	5.6
辽	43.4	13.9	6.2	7	6	11.2	8.3	4.1
吉	42.7	13.4	5.5	6	6	12.6	9.8	4
黑	40.5	14.7	6.1	8	6.5	10.8	9.1	4.4
苏	44.1	9	11.4	4.2	6	11.7	8.6	5
浙	40.3	8.5	10.6	6.7	7.9	12.2	8.8	5

续下页

城镇居民消费支出结构(续表)

地区	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
皖	47.3	11	7	3.2	6.4	13.2	8	3.9
闽	51.4	8.1	6.3	3.1	7.7	8.8	10.2	4.4
赣	44.9	8.7	6.7	3.1	6	11.3	14.6	4.6
鲁	37.1	13.6	12.2	4.9	6	13.3	8.2	4.7
豫	40.8	12.3	8.3	6	6.2	9.7	12	4.7
鄂	41.1	11.8	6.5	4.6	5.5	14.2	12.1	4.2
湘	40.5	10.7	8.4	4.3	6.7	14.5	10.3	4.7
粤	40.6	4.7	7.5	4.7	10.8	11.6	14.4	5.6
桂	44.3	6.6	7.4	3.4	7.2	13.6	12.8	4.8
琼	51.2	4.6	5	4.3	8.2	11.9	7.8	6.9
渝	42.3	10.8	9.5	4.3	7.4	13.4	8.1	4.1
川	43.9	11.3	7.7	4.5	5.3	12.8	9.6	5
黔	42.2	11	11.6	3.9	6.4	11.2	8.7	4.8
滇	44.4	10.9	7.5	5.1	5.9	11.4	8.3	6.7
藏	49.9	15.8	3.9	3.9	7.1	7	5.1	7.3
陕	37.3	9.9	11.3	6.6	5.8	12.4	11.9	4.8
甘	41.4	12.8	8.9	6	5.6	12.2	6.8	6.2
青	42.4	11.2	6.6	7.8	6.3	12.3	7.4	6
宁	38.8	13.6	7.7	8.9	7.1	12	6.4	5.5
新	38.6	12.9	10.4	5.7	6	13	8.3	5.1

10.12 在研究读写汉字能力与数学的关系的研究时,人们取得了232个美国亚裔学生的数学成绩和汉字读写能力的数据.关于汉字读写能力的变量有三个水平:“纯汉字”意味着可以完全自由使用纯汉字读写,“半汉字”意味着读写中只有部分汉字(比如日文),而“纯英文”意味着只能够读写英文而不会汉字.而数学成绩有4个水平(A、B、C、F).这里只选取亚裔学生是为了消除文化差

异所造成的影响. 这项研究是为了考察汉字具有的抽象图形符号的特性能否会促进儿童空间和抽象思维能力. 列联表形式数据如表10.4.

图 10.4 读写汉字能力与数学的关系数据

		数学成绩				总分
		数学A	数学B	数学C	数学F	
汉字	纯汉字	47	31	2	1	81
使用	半汉字	22	32	21	10	85
	纯英文	10	11	25	20	66
总分		79	74	48	31	231

表 10.6 服装定型的分类问题数据

	\mathbf{X}_1	\mathbf{X}_2	\mathbf{X}_3	\mathbf{X}_4	\mathbf{X}_5	\mathbf{X}_6	\mathbf{X}_7	\mathbf{X}_8	\mathbf{X}_9	\mathbf{X}_{10}	\mathbf{X}_{11}	\mathbf{X}_{12}	\mathbf{X}_{13}	\mathbf{X}_{14}	\mathbf{X}_{15}	\mathbf{X}_{16}
Xy_1	1.00															
\mathbf{X}_2	0.79	1.00														
\mathbf{X}_3	0.36	0.31	1.00													
Xy_4	0.96	0.74	0.38	1.00												
\mathbf{X}_5	0.89	0.58	0.31	0.90	1.00											
\mathbf{X}_6	0.79	0.58	0.30	0.78	0.79	1.00										
Xy_7	0.76	0.55	0.35	0.75	0.74	0.73	1.00									
\mathbf{X}_8	0.26	0.19	0.58	0.25	0.25	0.18	0.24	1.00								
\mathbf{X}_9	0.21	0.07	0.28	0.20	0.18	0.18	0.29	-0.04	1.00							
Xy_{10}	0.26	0.16	0.33	0.22	0.23	0.23	0.25	0.49	-0.34	1.00						
\mathbf{X}_{11}	0.07	0.21	0.38	0.08	-0.02	0.00	0.10	0.44	-0.16	0.23	1.00					
Xy_{12}	0.52	0.41	0.35	0.53	0.48	0.38	0.44	0.30	-0.05	0.50	0.24	1.00				
\mathbf{X}_{13}	0.77	0.47	0.41	0.79	0.79	0.69	0.67	0.32	0.23	0.31	0.10	0.62	1.00			
\mathbf{X}_{14}	0.25	0.17	0.64	0.27	0.27	0.14	0.16	0.51	0.21	0.15	0.31	0.17	0.26	1.00		
\mathbf{X}_{15}	0.51	0.35	0.58	0.57	0.51	0.26	0.38	0.51	0.15	0.29	0.28	0.41	0.50	0.63	1.00	
\mathbf{X}_{16}	0.21	0.16	0.51	0.26	0.23	0.00	0.12	0.38	0.18	0.14	0.31	0.18	0.24	0.50	0.65	1.00

第十一章 贝叶斯统计分析

本章概要

- ◇ 贝叶斯统计分析介绍
- ◇ 单参数与多参数贝叶斯分析
- ◇ 分层贝叶斯分析
- ◇ 线性回归与贝叶斯分析

§11.1 贝叶斯统计分析与经典统计比较

统计分析(推断)就是根据来自未知概率分布的观测数据, 对此分布或其参数作出推断, 如给出分布中参数的点估计、区间估计或对某假设进行检验. 统计分析主要有两大学派, 即贝叶学派与频率学派, 有时称为贝叶斯统计分析和经典统计分析, 在历史上它们之间曾产生过很大的分歧, 因为两者在统计推断的基本理论和方法上存在很大的差异. 但是, 现在它们之间开始相互尊重, 由此推动了现代数理统计的发展和许多实际问题的解决. 在叙述与讨论贝叶斯分析之前我们先对两者之间的差异作一个简单的对比.

11.1.1 经典统计分析中存在的问题

假设检验中的 p 值

贝叶斯统计学家认为, 经典假设检验中 p 值的计算违反了似然法则, 原因在于其涉及的数据信息超出了观测结果本身. 例如: 设随机变量 $Y \sim \text{Bin}(100, \theta)$,

其中 θ 为未知参数, 试验结果为 $y = 8$. 考虑假设检验问题

$$H_0 : \theta = 0.03 \leftrightarrow H_1 : \theta > 0.03$$

按经典的统计分析方法, 该检验的 p 值为

$$Pr(Y \geq y|\theta) = Pr(Y = 8|\theta = 0.03) + \cdots + Pr(Y = 100|\theta = 0.03).$$

可见, p 值的计算运用了大于等于观测结果的所有可能值. 而贝叶斯学派在解决此问题时, 则着重计算 $Pr(\theta > 0.03|Y = 8)$ 的值(称为 $\theta > 0.03$ 的后验概率).

置信区间

频率学派对于置信区间的解释是: 给定置信水平 $(1 - \alpha)$, 一个参数 θ 的 $100(1 - \alpha)\%$ 的置信区间就是一个按某种方法(如极大似然方法)构造的区间. 如果我们将试验重复多次, 并按这种方法计算出置信区间, 那么其中大约有 $100(1 - \alpha)\%$ 的比例包含参数 θ 的真值.

容易看出, 这种区间估计存在这样的问题:

- 1) 对于经典的统计学家而言, 参数 θ 是固定而未知的, 它没有分布可言. 因此频率学派不能说“有95%的概率使得参数 θ 落在置信区间中”.
- 2) 经典统计推断的精度主要取决于样本量的大小, 因而当数据量较少时, 该推断方法便难以实现.
- 3) 对于非对称分布、多峰分布及不可重复的数据, 此类置信区间很难获得.

造成这些问题的原因在于, 经典统计学派一贯用频率来解释概率, 并在此基础上理解一切统计推断的结论.

相反地, 贝叶斯学派认为参数是服从于某一分布的随机变量, 并结合数据信息与先验信息构造置信区间(在贝叶斯分析中通常称为可信区间), 使该未知参数以某一特定概率落入该区间. 这样, 贝叶斯学派可以说频率学派不可说的话: “有95%的概率这个区间包含参数 θ .”

11.1.2 对贝叶斯统计分析的质疑及褒奖

关于主观性

鉴于贝叶斯统计分析对于经典统计分析的强烈冲击, 其独树一帜的推断理

念也不断受到质疑. 首先, 在样本量较小时, 未知参数的估计往往对先验分布的选择相当敏感, 因而不少人认为此法过于主观. 其次, “将未知参数视为随机变量”这一想法也很难被接受.

贝叶斯统计学家对这两点质疑的回应则是, 在分析时可以通过模糊先验(diffuse prior)以及敏感性分析等方法弱化先验分布对结果的影响; 同时, 未知参数的随机性仅旨在体现对此参数所包含信息的一种不确定, 而非实际意义上的随机.

贝叶斯统计分析与经典统计分析的根本区别在于, 贝叶斯学派将已有的认识或知识视为是主观的(用先验分布来表示), 因而有时也被认为具有主观性. 但在大多数问题中, 鉴于其分析所用的模型囊括了似然函数以及先验分布两部分信息, 因此仍不失其客观性.

贝叶斯统计分析的优势

贝叶斯统计分析有着经典统计分析所无可比拟的优势, 主要有

- 1) 它结合了数据的信息与参数的先验信息, 不断通过样本数据更新先前的认知;
- 2) 与经典统计分析相比, 它的理论框架相对简洁, 且不需要繁杂的假设及数学推导;
- 3) 它不但能对缺失数据、截尾数据等进行简明处理, 还能对模型进行全面而稳健的估计.

事实上, 在不少统计问题中, 诸如线性回归、非参数统计等等, 经典统计方法仅仅是贝叶斯统计方法的特例. 同时, 贝叶斯统计分析还能直观地解释某些经典统计方法所无法阐述的问题, 诸如此前提到的置信区间问题等.

§11.2 贝叶斯统计分析与先验分布的选取

贝叶斯数据分析涉及两类估计量: 一类是通常所说的参数, 它(们)不可直接观测; 另一类则是可以潜在观测的量, 通常是待预测的未来的观测值. 贝叶斯统计分析的基本步骤包括:

- 1) 建立一个完整的概率模型. 它包括两部分, 即参数的先验分布和观测数据的抽样分布;

- 2) 对数据进行条件化得到后验分布: 计算后验分布, 并对它进行合理的解释;
- 3) 对模型的拟合及后验结果进行评估(包括合理性、敏感性等).

11.2.1 贝叶斯公式

贝叶斯学派的起点是贝叶斯的两项工作: 贝叶斯定理和贝叶斯假设. 贝叶斯定理(或贝叶斯公式)有三种形式.

贝叶斯公式的事件形式

设事件 A_1, A_2, \dots, A_k 为互不相容的事件, 它们的和包含事件 B , 即 $B \subset \bigcup_{i=1}^k A_i$, 则有

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{i=1}^k P(A_i)P(B|A_i)}, i = 1, 2, \dots, k. \quad (11-2.1)$$

贝叶斯公式的离散分布形式

设 X, Y 为随机变量, 其中 X 为离散型的, 其分布列为 $P(X = x_i) = p_X(x_i), i = 1, 2, \dots$. 当 $X = x_i$ 时, Y 对 X 的条件密度函数(若 Y 是连续的)或分布律(若 Y 为离散的)为 $P_{Y|X}(y|x)$, 则给定 $Y = y$ 时 X 对 Y 的条件列 $p_{X|Y}(x_i|y)$ 可表示为

$$p_{X|Y}(x_i|y) = \frac{p_X(x_i)p_{Y|X}(y|x_i)}{\sum_{j=1}^{\infty} p_X(x_j)p_{Y|X}(y|x_j)}, i = 1, 2, \dots \quad (11-2.2)$$

贝叶斯公式的连续分布形式

设随机变量 X, Y 的联合密度函数为 $p(x, y) = p_X(x)p_{Y|X}(y|x)$. 其中 $p_X(x)$ 为 X 的边际密度函数, $p_{Y|X}(y|x)$ 为当 $X = x$ 时 Y 对 X 的条件密度函数. 于是当 $Y = y$ 时 X 对 Y 的条件密度函数 $p_{X|Y}(x|y)$ 可表示为

$$p_{X|Y}(x|y) = \frac{p_X(x)p_{Y|X}(y|x)}{\int_{-\infty}^{\infty} p_X(x)p_{Y|X}(y|x)dx}. \quad (11-2.3)$$

贝叶斯定理

以后用 $y = (y_1, y_2, \dots, y_n)$ 表示数据, θ 表示不可观测的未知参数, 它可以是一维的, 也可以是多维的, \tilde{y} 表示一个未知但可潜在观测的量(简称为预测量). 在贝叶斯分析中 y 的抽样分布(也称为似然函数)表示为 $p(y|\theta)$. 在抽样之前, 我们对于 θ 可能有一定的了解(称为先验信息), 并用分布 $p(\theta)$ 来表示, 称之为 θ 的先验分布. 在没有样本信息时, 人们只能根据先验信息对 θ 作出推断; 在有了观测数据 y 后, 就可结合样本的信息与先验信息对 θ 作出推断, 而样本的信息与先验信息可以用 θ 的后验分布 $p(\theta|y)$ 进行综合, 在此基础上可以得到 \tilde{y} 的预测分布 $p(\tilde{y}|y)$.

由上面的贝叶斯公式可得:

1) 若 θ 是连续的, 则 θ 的后验分布可表示为

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{\int_{\Theta} p(\theta)p(y|\theta)d\theta} \quad (11-2.4)$$

2) 若 θ 是离散的, 则 θ 的后验分布可表示为

$$p(\theta_i|y) = \frac{p(\theta_i)p(y|\theta_i)}{\sum_j \pi(\theta_j)p(y|\theta_j)}, i = 1, 2, \dots \quad (11-2.5)$$

称这两个公式为**贝叶斯定理**(也称为贝叶斯公式、贝叶斯法则), 其中 $p(y) = \int_{\Theta} p(\theta)p(y|\theta)d\theta$ (连续场合)或 $p(y) = \sum_j \pi(\theta_j)p(y|\theta_j)$ (离散场合)为 y 的边际分布, 它们与参数 θ 无关. 以后若不作特别说明, 我们仅讨论参数是连续的场合. 由于 $p(y)$ 不含 θ 的任何信息, 因此用分布的核来表示, 贝叶斯定理可简化为

$$p(\theta|y) \propto p(\theta)p(y|\theta) \quad (11-2.6)$$

这是贝叶斯统计分析中常使用的贝叶斯公式的密度函数形式.

在 y 被观测到之前, 它是有分布可言的, 并称

$$p(y) = \int_{\Theta} p(\theta)p(y|\theta)d\theta$$

为 y 的边际分布或**先验预测分布**. 而当 y 一经观测得到, 我们就可对任一未知但

可观测的量 \tilde{y} 进行预测, 其后验分布为

$$\begin{aligned} p(\tilde{y}|y) &= \int_{\Theta} p(\tilde{y}, \theta|y) d\theta \\ &= \int_{\Theta} p(\tilde{y}|\theta, y) p(\theta|y) d\theta \\ &= \int_{\Theta} p(\tilde{y}|\theta) p(\theta|y) d\theta \quad (\text{因为 } y \text{ 与 } \tilde{y} \text{ 独立}). \end{aligned} \quad (11-2.7)$$

称之为 y 的后验预测分布.

11.2.2 先验分布的选取

贝叶斯统计中要使用先验信息, 而先验信息主要是指经验和历史资料. 因此如何利用人们的经验和过去的历史资料确定概率和先验分布是贝叶斯统计推断中一个关键性问题. 若人们已经获得有关参数的先验信息, 则可先确定先验密度函数, 然后根据专家的经验或利用经典的矩估计或极大似然估计确定先验分布中的参数(称为超参数). 先验密度函数的形式应有利于后验推断, 如选择共轭的先验分布. 若没有先验信息, 则可使用无信息先验分布.

为方便讨论, 先用贝叶斯观点叙述充分统计量的概念:

定义 11.2.1 设 y_1, y_2, \dots, y_n 表示来自总体 $p(y|\theta)$ 的样本. 对于参数 θ 而言, 统计量 $t(y_1, y_2, \dots, y_n)$ 称为充分的, 如果不论 θ 的先验分布是什么, 相应的后验分布 $p(\theta|y_1, y_2, \dots, y_n)$ 总是 θ 和 $t(y_1, y_2, \dots, y_n)$ 的函数.

定义11.2.1告诉我们, 后验分布是通过 $t(y_1, y_2, \dots, y_n)$ 与样本 y_1, y_2, \dots, y_n 发生联系的. 充分性的判定可使用著名的奈曼因子分解定理.

定理 11.2.1 若样本 y_1, y_2, \dots, y_n 对参数 θ 的条件密度 $p(y_1, y_2, \dots, y_n|\theta)$ 能表示成 $f(\theta, t(y_1, y_2, \dots, y_n))$ 与 $g(y_1, y_2, \dots, y_n)$ 的乘积, 则 $t(y_1, y_2, \dots, y_n)$ 对参数 θ 是充分的.

例 11.2.1 设 y_1, y_2, \dots, y_n 为来自正态总体 $N(\mu, 1)$ 的样本, 则样本均值 \bar{y} 是参数 μ 的充分统计量.

例 11.2.2 设 y_1, y_2, \dots, y_n 为来自正态总体 $N(0, \sigma^2)$ 的样本, 则样本观测值的平方和 $s^2 = \sum_{i=1}^n y_i^2$ 是参数 σ^2 的充分统计量.

先验分布的选取主要有三种方式:

1) 使用贝叶斯假设确定先验分布:

贝叶斯假设表述为: 参数 θ 的无信息先验分布 $p(\theta)$ 应在 θ 的取值范围 Θ 内是“均匀”分布, 用数学公式表示为 $p(\theta) = c, \theta \in \Theta$, 或 $p(\theta) \propto 1, \theta \in \Theta$, 其中 c 为常数, Θ 可为无限区间.

在贝叶斯假设下, 似然函数 $L(\theta|y_1, y_2, \dots, y_n)$ 为后验密度的核, 即

$$\pi(\theta|y_1, y_2, \dots, y_n) \propto L(\theta|y_1, y_2, \dots, y_n). \quad (11-2.8)$$

如果 $t(y_1, y_2, \dots, y_n)$ 为 θ 的充分统计量, 则上式可写成

$$p(\theta|t) \propto L(\theta|t). \quad (11-2.9)$$

尽管 $\pi(\theta) \propto 1, \theta \in \Theta$ 并不是正常的密度函数(有时称为广义密度函数), 而其后验密度(11-2.8)和(11-2.9)通常为正常的密度函数. 因此有时也称(11-2.8)或(11-2.9)为贝叶斯假设.

2) 使用杰弗莱原则确定无先验信息:

贝叶斯假设中的一个矛盾是: 如果对参数 θ 选用均匀分布, 那么当 θ 的函数 $g(\theta)$ 作为参数时, 也应该选用均匀分布作为先验分布. 然而由 θ 遵从均匀分布这一前提, 往往导出 $g(\theta)$ 不是均匀分布, 反之亦然. 杰弗莱为了克服这一矛盾提出了选取先验的不变原理——并被称为杰弗莱原则.

杰弗莱原则有两个部分: 一是对无信息先验分布有一合理的要求; 另一部分是给出一个具体的方法去求得符合要求的先验分布. 现设按照同一准则决定的 θ 的先验分布为 $p(\theta)$, $\eta = g(\theta)$ 的先验分布为 $p_g(\eta)$, 由它们应满足关系:

$$p(\theta) = p_g(g(\theta))|g'(\theta)|. \quad (11-2.10)$$

杰弗莱巧妙地利用了费歇信息阵的一个不变性质, 找到了满足(11-2.10)要求的先验分布 $p(\theta)$: θ 的无信息先验分布应以信息阵 $I(\theta)$ 的行列式的平方根为核, 即

$$p(\theta) \propto |I(\theta)|^{1/2}, \quad (11-2.11)$$

其中 θ 可以是向量,

$$I(\theta) = E \left(\frac{\partial \ln p(x_1, \dots, x_n | \theta)}{\partial \theta} \right) \left(\frac{\partial \ln p(x_1, \dots, x_n | \theta)}{\partial \theta} \right)'. \quad (11-2.12)$$

由于 $I(\theta)$ 是非负定的, $|I(\theta)| > 0$, 因此 $|I(\theta)|^{1/2}$ 有意义. 按(11-2.11)所确定的先验分布的确具有不变性, 因为我们可以证明下面的定理.

定理 11.2.2 设 $g(\theta)$ 是 θ 的函数, $\eta = g(\theta)$ 与 θ 具有相同的维数, 则有

$$|I(\theta)|^{1/2} = \left| \frac{\partial g(\theta)}{\partial \theta} \right| |I(\eta)|^{1/2}. \quad (11-2.12)$$

3) 共轭分布法

定义 11.2.2 设 y_1, y_2, \dots, y_n 表示来自总体 $p(y|\theta)$ 的样本. 先验分布 $p(\theta)$ 称为 θ 的共轭先验分布, 如果后验分布 $p(\theta|y_1, y_2, \dots, y_n)$ 与 $p(\theta)$ 是同一类型的, 即它们的核有相同的形式.

共轭先验分布的二个优点: 1) 计算方便; 2) 后验分布的一些参数可以得到很好的解释(见后面的例子). 然而贝叶斯统计中先验分布的选取以合理性作为首要原则, 而计算的方便是第二位的. 一般的做法是: 在没有具有的先验信息时, 采用贝叶斯假设或更为一般的杰弗莱原则采用无信息先验分布. 但是使用贝叶斯统计分析方法的主要目的是充分利用专家的经验 and 历史数据, 特别是在小样本场合与多参数场合经典的统计分析方法显得特别困难或无能为力时, 这时选取一个合理的先验分布. 先验分布选取的合理性显得尤为重要: 贝叶斯统计分析中先验分布的选取带有主观性(这是这种方法受到批评或攻击的原因), 我们使用贝叶斯统计分析方法应尽可能将先验信息通过先验分布客观地反映到统计分析中, 以弥补数据中信息的不足, 从而达到客观合理地解决实际问题的目的. 具体方法可参考茆诗松的教材(1999).

11.2.3 贝叶斯分析体现了科学探索过程

设已观测到数据 y_1 , 则由贝叶斯定理, 得后验分布

$$p(\theta|y_1) \propto p(y_1|\theta) \times p(\theta).$$

假设后来又观测到数据 y_2 (与 y_1 独立), 则

$$p(y_1, y_2|\theta) = p(y_1|\theta) \times p(y_2|\theta).$$

因此, 再由贝叶斯定理, 得后验分布

$$p(\theta|y_1, y_2) \propto p(\theta) \times p(y_1|\theta) \times p(y_2|\theta)$$

$$= p(\theta|y_1) \times p(y_2|\theta).$$

从上述公式我们可以看到这样一个过程：由数据 y_1 对先验 $p(\theta)$ 作出更新得到后验分布 $p(\theta|y_1)$ ；在观测到数据 y_2 后，将 $p(\theta|y_1)$ 视为新的先验分布，并由 y_2 对它作出更新得到后验分布 $p(\theta|y_1, y_2)$ 。这个过程可以不断重复，由此随着对于参数 θ （或相关的分布）的信息不断增加，将得出更符合实际（数据）的结论。这个过程体现了科学研究的不断探索过程。

注：本章随机变量或其观测值均用小写字母表示。在试验或观测之前它被视为是随机的，其分布称为抽样分布或数据分布；在试验或观测之后它就被视为样本或数据，这时抽样分布就转化为似然函数（作为参数的函数）。

§11.3 单参数贝叶斯统计分析

所谓单参数模型（分布），即统计模型中仅含有一个未知参数。常用的单参数模型有两项分布、正态分布（其中仅有一个参数未知）、泊松分布、指数分布等。本节讨论基于此类模型下的贝叶斯统计推断。

11.3.1 两项分布下的贝叶斯推断

设随机变量 y 代表 n 次贝努利试验中的某事件A“成功”的次数，参数 θ 代表每次试验成功（事件A发生）的概率。 θ 也可表示总体中具有某种特征的个体所占的比例。由于 n 次贝努利试验独立，因此 y 服从二项分布 $\text{Bin}(n, \theta)$ ，即

$$p(y|\theta) = \text{Bin}(y|n, \theta) = C_n^y \theta^y (1 - \theta)^{(n-y)}. \quad (11-3.1)$$

在两项分布模型下进行贝叶斯统计推断，还需给出参数 θ 的先验分布。

1) 基于贝叶斯假设的无信息先验分布

若事先没有关于参数 θ 的任何信息，则通常假设其服从 $[0, 1]$ 上的均匀分布，即得到最简单的先验分布 $p(\theta) \propto 1$ ，即贝塔分布 $\text{Beta}(1, 1)$ 。从而根据贝叶斯公式导出 θ 的后验分布为：

$$p(\theta|y) \propto \theta^y (1 - \theta)^{(n-y)},$$

即

$$\theta|y \sim \text{Beta}(y + 1, n - y + 1). \quad (11-3.2)$$

2) 基于杰弗莱原则的无信息先验分布

由(11-3.1)不难得到(请读者自己证)

$$I(\theta) = \frac{n}{\theta(1-\theta)},$$

因此由杰弗莱原则, 得到 θ 的先验分布

$$p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2},$$

即 θ 服从贝塔分布 $\text{Beta}(1/2, 1/2)$. 因此由贝叶斯公式得到 θ 的后验分布为:

$$p(\theta|y) \propto \theta^{y-1/2}(1-\theta)^{n-y-1/2},$$

即

$$\theta|y \sim \text{Beta}(y+1/2, n-y+1/2). \quad (11-3.3)$$

3) 基于共轭先验分布

取 θ 的先验分布为贝塔分布 $\text{Beta}(\alpha, \beta)$, 即

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

由抽样分布(11-3.1)及先验分布得 θ 的后验分布为:

$$p(\theta|y) \propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1},$$

即

$$\theta|y \sim \text{Beta}(y+\alpha, n-y+\beta). \quad (11-3.4)$$

由此我们可以得到结论:

- 1) 对于二项分布中参数 θ , 基于贝叶斯假设与基于杰弗莱原则的先验可以视为共轭先验的特例;
- 2) 由于贝塔分布 $\text{Beta}(\alpha, \beta)$ 分布的均值与方差分别为

$$E(\theta) = \frac{\alpha}{\alpha+\beta}, \quad \text{Var}(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)},$$

故在共轭先验下参数 θ 的贝叶斯估计(后验均值)为

$$\hat{\theta} = \frac{y + \alpha}{n + \alpha + \beta}. \quad (11-3.5)$$

- 3) θ 的贝叶斯估计的解释: $\frac{\alpha}{\alpha+\beta}$ 可视为仅利用先验信息对参数 θ 的估计, $\frac{y}{n}$ 则为利用样本对参数 θ 的估计, 而 θ 的后验贝叶斯估计(11-3.5)为上述二个估计的加权平均, 即

$$\frac{y + \alpha}{n + \alpha + \beta} = \frac{\alpha + \beta}{n + \alpha + \beta} \times \frac{\alpha}{\alpha + \beta} + \frac{n}{n + \alpha + \beta} \times \frac{y}{n}$$

- 4) 一个新的观测值 \tilde{y} (假定与 y_1, \dots, y_n 独立)的预测值为

$$\begin{aligned} Pr(\tilde{y} = 1|y) &= \int_0^1 Pr(\tilde{y} = 1|\theta, y)p(\theta|y)d\theta \\ &= \int_0^1 \theta p(\theta|y)d\theta = E(\theta|y) = \frac{y + \alpha}{n + \alpha + \beta}. \end{aligned}$$

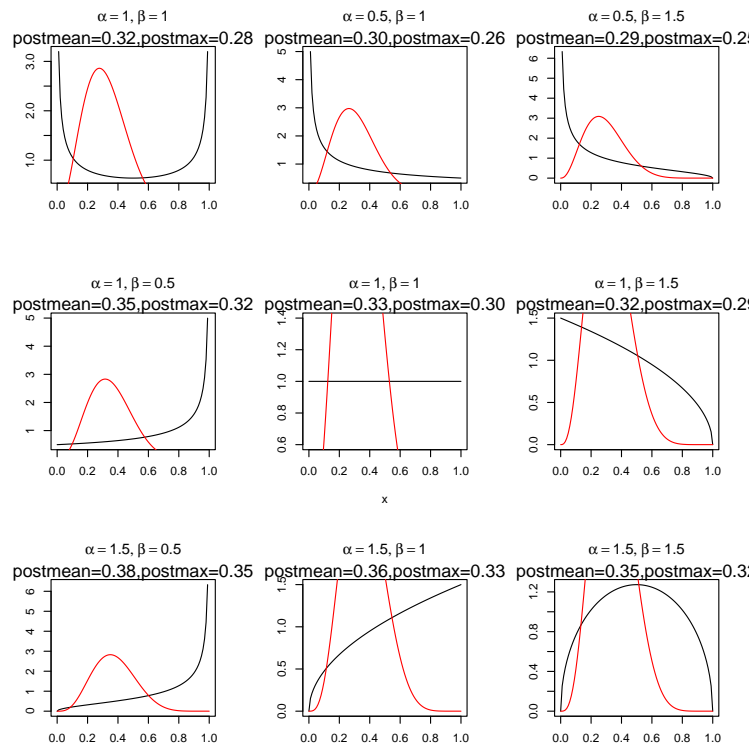
下面通过一些例子来说明贝叶斯分析中的影响因素.

例 11.3.1 相同数据下, 不同先验分布对贝叶斯分析的影响: 设 $y|\theta \sim Bin(n, \theta)$, $n = 10, y = 3$, 先验分布取为 $Beta(\alpha, \beta)$, 其中超参数取9组 $(\alpha, \beta) = (0.5, 0.5), (0.5, 1.0), (0.5, 1.5), (1.0, 0.5), (1.0, 1.0), (1.0, 1.5), (1.5, 0.5), (1.5, 1.0), (1.5, 1.5)$. 图11.1给出了9种先验分布及相应的后验分布(R程序从略).

结论: 在相同的数据下, 先验对于后验有一定的影响, 但这种影响不是很明显.

例 11.3.2 相同先验及样本容量下, 不同观测对贝叶斯分析的影响: 设 $y|\theta \sim Bin(n, \theta)$, $n = 10$, θ 取无信息先验 $Beta(1, 1)$, 图11.2给出了 $y = 0, 1, 2, 3, 4, 5$ 时6种后验分布. R程序如下:

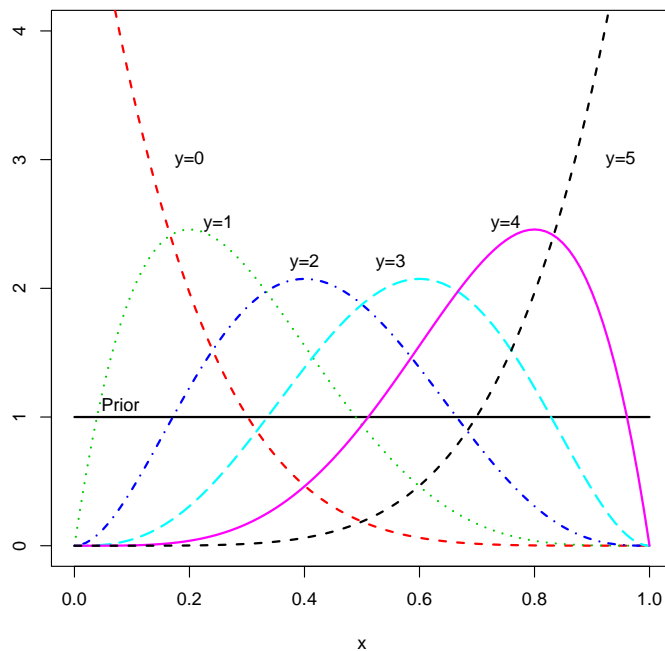
```
> x=seq(0,1,0.01)
> n=5
> z=dbeta(x,1,1)
> z0=dbeta(x,1,n+1)
> z1=dbeta(x,1+1,n-1+1)
> z2=dbeta(x,2+1,n-2+1)
> z3=dbeta(x,3+1,n-3+1)
```

图 11.1 后验分布: $\mathcal{B}e(\alpha + y, n - y + \beta)$

```

> z4=dbeta(x,4+1,n-4+1)
> z5=dbeta(x,5+1,n-5+1)
> z.df=data.frame(cbind(z,z0,z1,z2,z3,z4,z5))
> matplot(x, z.df, ylim=c(0,4), xlab="x", ylab="",
          col=1:6, type="l", lwd=2)
> text(0.20,3,"y=0")
> text(0.25,2.5,"y=1")
> text(0.4,2.2,"y=2")
> text(0.55,2.2,"y=3")
> text(0.75,2.5,"y=4")
> text(0.95,3,"y=5")
> text(0.08,1.1,"Prior")

```

图 11.2 后验分布: $Be(\alpha + y, n - y + \beta)$

结论: 在相同的样本容量下, 不同的观测对后验的影响很明显.

例 11.3.3 随着观测信息的增加(样本容量的增加), 不同先验对贝叶斯分析的影响: 设 $y|\theta \sim Bin(n, \theta)$, θ 取3种先验分布 $Beta(1, 1)$, $Beta(2, 5)$, $Beta(10, 1)$, 考查2种观测数据: $n = 5, y = 1$ 和 $n = 50, y = 10$, 这时经典的极大似然估计均为 $\hat{\theta} = 0.2$. 图11.3给出了3种先验对后验分布的影响与样本容量的关系. **R**程序如下:

```
> x=seq(0,1,0.01)
> par(mfrow=c(1,3))
> # 左侧图形 -- 先验
> z1=dbeta(x,1,1);
> z2=dbeta(x,5,2);
> z3=dbeta(x,1,10)
> z.df=data.frame(cbind(z1,z2,z3))
```

```
> matplot(x,z.df, xlab="y", ylab="",
          col=c("black", "red", "blue"),
          type="l", lty=1:3, lwd=2)
> text(0.8,6,"Priors")
> # 中间图形 -- 后验: n=5, y=1
> n=5
> y=1
> z1=dbeta(x,y+1,n-y+1);
> z2=dbeta(x,y+5,n-y+2);
> z3=dbeta(x,y+1,n-y+10)
> z.df=data.frame(cbind(z1,z2,z3))
> matplot(x,z.df, xlab="y", ylab="",
          col=c("black", "red", "blue"),
          type="l", lty=1:3, lwd=2)
> text(0.8,3.5,"Posterios")
> text(0.8,3.3,"n=5, y=1")
> # 右侧图形 -- 后验: n=50, y=10
> n=50
> y=10
> z1=dbeta(x,y+1,n-y+1);
> z2=dbeta(x,y+5,n-y+2);
> z3=dbeta(x,y+1,n-y+10)
> z.df=data.frame(cbind(z1,z2,z3))
> matplot(x,z.df, xlab="y", ylab="",
          col=c("black", "red", "blue"),
          type="l", lty=1:3, lwd=2)
> text(0.8,4.5,"Posterios")
> text(0.8,4.3,"n=50, y=10")
```

结论: 随机样本容量的增加, 先验对后验的影响逐渐减小. 这说明在小样本场合, 先验的选取较为重要, 但随机数据信息的增加, 先验在贝叶斯分析中的敏感性较弱, 其选择可以考虑以方便计算为主, 如共轭先验.

4) 基于后验分布的推断

前面我们已经看到后验分布是已有的数据信息对先验信息更新调整的结

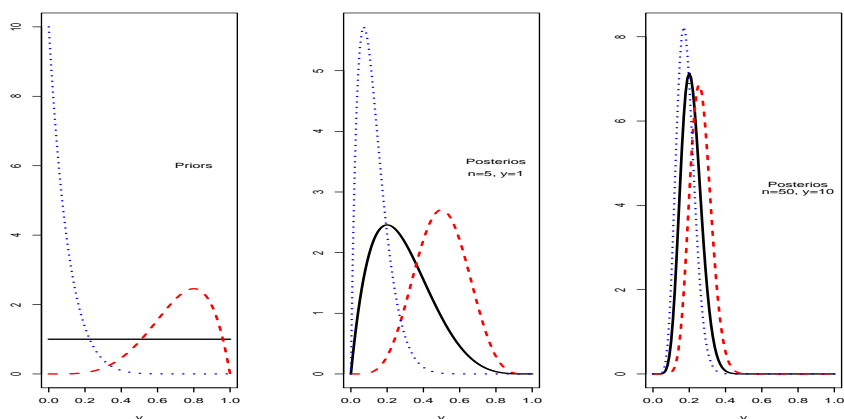


图 11.3 随机样本信息的增加的后验分布变化

果,它概括了参数的一切信息,如我们最为关心的后验均值、后验众数、后验方差或标准差、后验置信区间.后验均值、后验众数都可作为贝叶斯点估计,但前者更为常用,在后验分布对称或近似对称时,二者一致;后验方差或标准差反映了贝叶斯估计(点估计或区间估计)的精度;贝叶斯置信区间有两种形式,其一是等尾的置信区间,这与经典的区间估计一致,其二称为最高概率密度(Highest Probability Density)置信区间,其优点是在相同的置信水平下这样的区间估计是最短的,但其计算需要使用数值方法.本书讲的贝叶斯置信区间是指等尾的,它很容易由后验分布的分位数求得.

另外,在一些实际问题中,我们不仅关心分布中的参数本身,更关心其函数的统计性质.例如,在考虑某地区女性出生率时,样本可用二项分布 $\text{Bin}(n, \theta)$ 描述,这时我们不仅关心参数 θ ,更为关心的是男性与女性出生的比率 $\phi = \frac{1-\theta}{\theta}$,有时在研究社会问题时还关心 θ 的logit变换 $\text{logit}(\theta) = \log(\theta/(1-\theta))$.对于 θ 的变换,其后验分布通常不易获得,但我们可以通过从 θ 的后验分布中随机抽取一系列的样本,从而获得其函数的样本,当抽样次数足够大时,理论上可以获得这些参数或其函数的精确后验分布,因而我们只要基于后验样本进行推断就足够了.这是贝叶斯统计分析中最为常用的方法,在多参数场合其优势更为明显.另外,对参数进行适当的变换,可以使后验分布具有更好的对称性,这样可以借助正态近似求得它们的贝叶斯置信区间.下面通过一个具体的例子予以说明.

例 11.3.4 早期在德国进行了一项试验,其结果显示,在980例因非正常受孕而导致胎盘位置过低的分娩中,有437例为女婴.由此能否判断在此类非

正常分娩中,女婴的出生率小于0.485呢?

解 可以认为这980名孕妇中,女婴的出生数 y 服从二项分布 $\text{Bin}(n, \theta)$, 现已知 $n = 980, y = 437$. 假设我们对 θ 没有任何可用的信息, 故用无信息先验分布 $\text{Beta}(1, 1)$ 作为 θ 的先验. 则 θ 的后验分布为 $\theta|y \sim \text{Beta}(438, 544)$.

1) 首先计算 θ 后验均值、标准差、中位数及95%置信区间, **R**程序如下:

```
> alpha <- 438
> beta <- 544
> postmean<-alpha/(alpha+beta)
> print("The posterior mean is")
> print(postmean)
> poststd<-sqrt(alpha*beta/(alpha+beta)^2/(alpha+beta+1))
> print("The posterior standard deviation is")
> print(poststd)
> postmedian<-qbeta(0.5, alpha, beta)
> print("The median based on
> posterior distribution is")
> print(postmedian)
> CI_95<-c(qbeta(0.025,alpha, beta), qbeta(0.975, alpha, beta))
> print("The 95% posterior confidence interval is")
> print(CI_95)
```

结论: 运行上述程序得到 θ 后验均值、标准差、中位数及95%置信区间分别为:

```
0.4460285
0.01585434
0.4459919
(0.4150655 0.4771998)
```

2) 运用随机模拟方法, 根据 θ 的后验分布进行推断, 即产生1000个 $\text{Beta}(438, 544)$ 的随机数, 并计算其后验均值、标准差、中位数和基于正态近似的95%置信区间. **R**程序如下:

```
> alpha <- 438; beta <- 544
```

```

> theta <- rbeta(1000, alpha, beta)
> sort_theta <- sort(theta)
> spostmean <- mean(theta)
> spoststd <- sd(theta)
> spostmedian <- sum(sort_theta[500:501])/2
> approxCI_95 <- c(spostmean-1.96*spoststd, spostmean+1.96*spoststd)
> print(spostmean)
> print(spoststd)
> print(spostmedian)
> print("The 95% confidence interval of theta
        based on normal approximation is")
> print(approxCI_95)

```

结论: 运行上述程序得到 θ 后验均值、标准差、中位数及95%置信区间分别为:

```

0.4463512
0.01531046
0.4461134
(0.4163427, 0.4763597)

```

它们与直接从后验分布计算的结果几乎没有差异.

- 3) 基于随机模拟, 计算两种变换 $\text{logit}(\theta)$ 、 $\phi = (1 - \theta)/\theta$ 的后验均值、标准差、中位数和基于正态近似的95%置信区间. **R**程序如下:

```

> alpha <- 438; beta <- 544
> theta <- rbeta(1000, alpha, beta)
> logit_theta <- log(theta/(1-theta))
> sort_logit_theta <- sort(logit_theta)
> slogit_median <- sum(sort_logit_theta[500:501])/2
> slogit_postmean <- mean(logit_theta)
> slogit_poststd <- sd(logit_theta)
> L <- slogit_postmean-1.96*slogit_poststd
> U <- slogit_postmean+1.96*slogit_poststd
> approxlogit_CI = c(L, U)
> approx_CI = c(exp(L)/(1+exp(L)), exp(U)/(1+exp(U)))

```

```

> print(slogit_postmean)
> print(slogit_poststd)
> print(slogit_median)
> print("The 95% confidence interval of logit(theta)
      based on normal approximation is")
> print(approxlogit_CI)
> print("The 95% confidence interval of theta
      based on normal approximation is")
> print(approx_CI)
> # phi的推断
> phi<-(1-theta)/theta
> sort_phi <- sort(phi)
> sphi_median <- sum(sort_phi[500:501])/2
> sphi_postmean<-mean(phi)
> sphi_poststd <- sd(phi)
> L<-sphi_postmean-1.96*sphi_poststd
> U<-sphi_postmean+1.96*sphi_poststd
> approxphi_CI<-c(L, U)
> print(sphi_postmean)
> print(sphi_poststd)
> print(sphi_median)
> print("The 95% confidence interval of phi=(1-theta)/theta is" )
> print(approxphi_CI)

```

结论: 运行上述程序得到 $\text{logit}(\theta)$ 后验均值、标准差、中位数及95%置信区间分别为:

```

-0.2157638
 0.0635593
-0.2182796
(-0.34034002, -0.09118757)

```

而 $\phi = (1 - \theta)/\theta$ 后验均值、标准差、中位数及95%置信区间分别为:

```

1.243318
0.07920171

```

```
1.243803
(1.088082, 1.398553)
```

另外, 由 $\text{logit}(\theta)$ 的反变换得 θ 的95%置信区间为(0.4157269, 0.4772189), 与前面的结果也基本相同.

4) 作出 θ , $\text{logit}(\theta)$ 和 $\phi = (1 - \theta)/\theta$ 的频数直方图. **R**程序如下:

```
> alpha <- 438; beta <- 544
> theta <- rbeta(1000,alpha,beta)
> par(mfrow=c(1,3))
> #Fig(1,1)-- histogram of theta
> par(mar=c(5,4,2,1))
> hist(theta, breaks = seq(0.35,0.55,0.005),
      xlim = c(0.35,0.55),
      main="", xlab=quote(theta),
      probability="T")
> #Fig(1,2) -- histogram of log(theta)
> logit_theta <- log(theta/(1-theta))
> breaks <- quantile(logit_theta, 0:20/20)
> par(mar=c(5,4,2,1))
> hist(logit_theta, breaks = seq(-0.5,0.1,0.01),
      xlim = c(-0.5,0.1), main="",
      xlab=quote(logit(theta)==log(theta/(1-theta))),
      probability=T)
> #Fig(1,3) -- histogram of phi=(1-theta)/theta
> phi=(1-theta)/theta
> breaks <- quantile(phi, 0:20/20)
> par(mar=c(5,4,2,1))
> hist(phi, breaks = seq(0.8,1.6,0.01),
      xlim = c(1.0,1.6),
      main="", xlab=quote(phi==(1-theta)/theta),
      probability=T)
```

结论: 运行上述程序得到图11.4. 由于欧洲人种新生儿的男女比率一般为1.06(即女婴出生率为0.485), 因此根据 ϕ 的中位数及其基于正态近似下得95%置信区间(1.088082, 1.398553)推断, 女婴出生率在上述非正常分娩

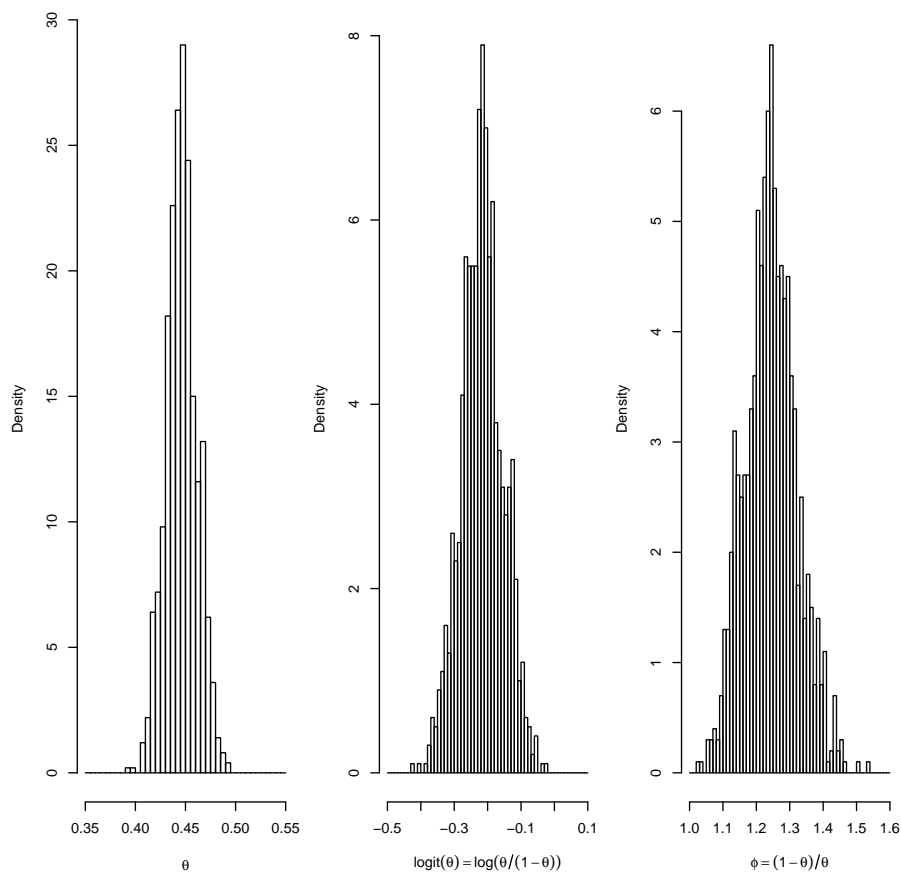


图 11.4 由 θ 的后验分布产生的1000个随机样本的直方图.

状况下, 确实比一般情况下要低. 此例也可利用共轭先验分布假设进行推断, 结果相同.



11.3.2 正态分布下的贝叶斯统计推断

我们先考虑仅有一个观测值的情形, 然后推到多个观测值的一般情形.

1) 单一正态观测值, 方差已知时

设观测值 y 服从正态分布 $N(\theta, \sigma^2)$, 其中 σ^2 已知, 则 y 的似然函数为

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\theta)^2}{2\sigma^2}\right).$$

根据似然函数的形式不难推得 θ 的共轭先验密度为正态分布 $\theta \sim N(\mu_0, \tau_0^2)$, 即

$$p(\theta) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right),$$

其中 μ_0, τ_0^2 为超参数(假定已知), 由此可得 θ 的后验密度

$$p(\theta|y) \propto \exp\left(-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta - \mu_0)^2}{\tau_0^2}\right]\right).$$

经整理后得

$$p(\theta|y) \propto \exp\left(-\frac{1}{2\tau_1^2}(\theta - \mu_1)^2\right), \quad (11-3.6)$$

即 $\theta \sim N(\mu_1, \tau_1^2)$, 其中

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}, \quad (11-3.7)$$

$$\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}. \quad (11-3.8)$$

由此我们得到结论:

- 在正态分布中, 方差的倒数称为**精度**. 因此(11-3.7)表示后验均值等先验均值与观测值 y 的加权平均, 其权数就是两者相应的精度.
- (11-3.8)后验精度等于先验精度与数据精度之和.

最后再考虑后验预测分布. 由(11-2.7)得, 未来观测值 \tilde{y} 的预测分布为

$$\begin{aligned} p(\tilde{y}|y) &= \int_{\Theta} p(\tilde{y}|\theta)p(\theta|y)d\theta \\ &\propto \int_{\Theta} \exp\left(-\frac{1}{2\sigma^2}(\tilde{y} - \theta)^2\right) \exp\left(-\frac{1}{2\tau_1^2}(\theta - \mu_1)^2\right) d\theta. \end{aligned}$$

由于上述积分的被积函是 (\tilde{y}, θ) 二次型的指数, 因此 (\tilde{y}, θ) 服从联合正态分布, 从而 \tilde{y} 的边缘分布, 即 $p(\tilde{y}|y)$ 是正态的. 因此我们只需求出其期望 $E(\tilde{y}|y)$ 和方差 $\text{Var}(\tilde{y}|y)$. 而由 $E(\tilde{y}|\theta) = \theta$, $\text{Var}(\tilde{y}|\theta) = \sigma^2$ 可得

$$E(\tilde{y}|y) = E(E(\tilde{y}|\theta, y)|y) = E(\theta|y) = \mu_1,$$

$$\begin{aligned}\text{Var}(\tilde{y}|y) &= E(\text{Var}(\tilde{y}|\theta, y)|y) + \text{Var}(E(\tilde{y}|\theta, y)|y) \\ &= E(\sigma^2|y) + \text{Var}(\theta|y) \\ &= \sigma^2 + \tau_1^2.\end{aligned}$$

所以

$$\tilde{y}|y \sim N(\mu_1, \sigma^2 + \tau_1^2).$$

由此我们得到结论:

- \tilde{y} 的预测分布的均值等于后验均值;
- 预测分布的方差等于模型的方差与来自 θ 的后验不确定性的方差 τ_1^2 之和.

2) 多个正态观测值, 方差已知时

设 $y = (y_1, \dots, y_n)$ 为一系列独立同分布的观测值, $y_i \sim N(\theta, \sigma^2), i = 1, \dots, n$, 其中 σ^2 已知. 则 $\bar{y} = \sum_{i=1}^n y_i/n$ 为充分统计量, 且

$$\bar{y}|\theta \sim N(\mu, \sigma^2/n).$$

由此不难将多观测值情形转化为单一观测值情形来研究, 从而得到

$$p(\theta|y_1, \dots, y_n) = p(\theta|\bar{y}) = N(\theta|\mu_n, \tau_n^2), \quad (11-3.9)$$

其中

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \quad (11-3.10)$$

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}. \quad (11-3.11)$$

未知观测值 \tilde{y} 的预测分布为

$$\tilde{y}|\bar{y} \sim N(\mu_n, \sigma^2 + \tau_n^2). \quad (11-3.12)$$

值得一提的是, 在先验精度较低或样本容量较大时(即 $\tau_0^2 \rightarrow \infty$, n 固定; 或 $n \rightarrow \infty$, τ_0^2 固定), 则有如下近似

$$p(\theta|y) \approx N(\theta|\bar{y}, \sigma^2/n).$$

此即无信息均匀先验 $p(\theta) \propto 1$ 下的后验分布, 这是容易理解的.

3) 多个正态观测值, 均值已知时

若 $y_1, \dots, y_n \stackrel{iid}{\sim} N(\theta, \sigma^2)$, 其中 θ 已知. 此时似然函数为

$$p(y|\sigma^2) \propto (\sigma^2)^{-n/2} \exp\left(-\frac{n}{2\sigma^2} s_\mu^2\right), \quad (11-3.13)$$

其中

$$s_\mu^2 = \frac{1}{2} \sum_{i=1}^n (y_i - \theta)^2$$

为充分统计量. 由于 σ^2 的共轭先验分布为逆伽玛分布 $\text{IGa}(\alpha, \beta)$, 即

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp(-\beta/\sigma^2),$$

在此先验下, σ^2 的后验分布为

$$p(\sigma^2|s_\mu^2) \propto (\sigma^2)^{-(\alpha + \frac{n}{2} + 1)} \exp\left(-\frac{\beta + s_\mu^2/2}{\sigma^2}\right),$$

即

$$\sigma^2|s_\mu^2 \sim \text{IGa}\left(\alpha + \frac{n}{2}, \beta + \frac{s_\mu^2}{2}\right). \quad (11-3.14)$$

4) 泊松分布

单位时间内、单位面积或单位空间中某事件(设为A)发生的次数常可用泊松分布来刻画, 例如单位时间内飞机起飞或下降的次数、单位时间内某交通路口通过的车辆数、单位面积内的害虫数等等都可用泊松分布来描述.

设 $y = (y_1, \dots, y_n)$ 为来自泊松分布 $\text{Poisson}(\theta)$ 的容量为 n 的样本, 参数 θ 表

示事件出现的频率(强度). 则 $y = (y_1, \dots, y_n)$ 的似然函数为:

$$\begin{aligned} L(\theta|y) &= p(y|\theta) = \prod_{i=1}^n \frac{1}{y_i!} \theta^{y_i} e^{-\theta} \\ &\propto \theta^{t(y)} e^{-n\theta}, \end{aligned} \quad (11-3.15)$$

其中 $t(y) = \sum_{i=1}^n y_i$ 为充分统计量. 由于 θ 的共轭先验分布为伽玛分布 $\text{Ga}(\alpha, \beta)$, 即

$$p(\theta) \propto e^{-\beta\theta} \theta^{\alpha-1}$$

与似然函数对照发现, 我们可以将先验理解为: 在 β 次先验观察中事件 A 出现了 $\alpha - 1$ 次.

由此容易得到 θ 的后验分布

$$\theta|y \sim \text{Gamma}(\alpha + n\bar{y}, \beta + n), \quad (11-3.16)$$

其中 $\bar{y} = t(y)/n$.

一个自然的推广是: $y_i \sim \text{Poisson}(x_i\theta), i = 1, 2, \dots, n$. 在流行病学研究中 θ 为某疾病的发病率, x_i 为个体 i 的暴露(可能受感染)时间, 这时有似然函数

$$L(\theta|y) = p(y|\theta) \propto \exp \left\{ \log(\theta) \sum_{i=1}^n y_i - \theta \sum_{i=1}^n x_i \right\}. \quad (11-3.17)$$

θ 仍取共轭先验 $\text{Ga}(\alpha, \beta)$, 则其后验分布为

$$\theta|y \sim \text{Gamma} \left(\alpha + \sum_{i=1}^n y_i, \beta + \sum_{i=1}^n x_i \right). \quad (11-3.18)$$

5) 指数分布

以一定的频率 θ 独立出现的事件的时间间隔(等待时间)可用指数分布来描述, 许多产品的寿命也可用指数来刻画, 其密度函数为

$$p(y|\theta) = \theta \exp\{-y\theta\}, \quad y > 0,$$

记为 $\text{Exp}(1/\theta)$. 指数分布具有独特的无记忆性, 即

$$Pr(y > t + s | y > s, \theta) = Pr(y > t | \theta) \quad \forall s, t > 0.$$

对于数据 $y = (y_1, \dots, y_n)$, $y_i \sim \text{Exp}(1/\theta)$, 基于共轭先验分布 $\text{Ga}(\alpha, \beta)$, 得 θ 的后验分布

$$\theta | y \sim \text{Gamma} \left(\alpha + n, \beta + \sum_{i=1}^n y_i \right). \quad (11-3.19)$$

§11.4 多参数贝叶斯统计分析

许多实际的统计问题都含有多个未知参数, 但人们通常只对其中的一部分参数感兴趣, 其余参数称为“讨厌”参数. 在处理这类问题时, 贝叶斯方法与其他传统的推断方法有明显的优势.

11.4.1 方法概述

假设参数(向量) θ 由两部分组成, $\theta = (\theta_1, \theta_2)$, 其中 θ_1 为感兴趣的参数, θ_2 为讨厌参数. 设数据 y 的分布为 $p(y|\theta_1, \theta_2)$, θ 的先验分布为 $p(\theta_1, \theta_2)$, 则 θ_1 与 θ_2 联合后验密度函数为

$$p(\theta_1, \theta_2 | y) \propto p(y | \theta_1, \theta_2) p(\theta_1, \theta_2). \quad (11-4.1)$$

在联合后验密度函数中对 θ_2 求积分, 得到 θ_1 的边际后验密度

$$\begin{aligned} p(\theta_1 | y) &= \int p(\theta_1, \theta_2 | y) d\theta_2 \\ &= \int p(\theta_1 | \theta_2, y) p(\theta_2 | y) d\theta_2 \end{aligned} \quad (11-4.2)$$

11.4.2 正态分布参数中的贝叶斯分析

设 $y = (y_1, \dots, y_n) \stackrel{iid}{\sim} N(\mu, \sigma^2)$, 其中 μ 和 σ^2 均未知. 在此仅考虑独立无信息先验

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}.$$

此时 (μ, σ^2) 的联合后验密度为

$$\begin{aligned} p(\mu, \sigma^2 | y) &\propto \sigma^{-n-2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right\} \\ &= \sigma^{-n-2} \exp \left\{ -\frac{1}{2\sigma^2} [(n-1)s^2 + n(\bar{y} - \mu)^2] \right\} \end{aligned}$$

其中样本均值 \bar{y} 与样本方差 s^2 为充分统计量. 上式对 μ 求积后得到

$$p(\sigma^2 | \bar{y}, s^2) \propto \sigma^{-(n+1)/2} \exp \left\{ -\frac{1}{2\sigma^2} [(n-1)s^2] \right\}, \quad (11-4.3)$$

即 σ^2 的后验服从倒伽玛分布 $IGa(\frac{n-1}{2}, \frac{(n-1)s^2}{2})$.

在实际问题中总体均值 μ 通常是感兴趣的参数. 将联合后验密度对 σ^2 求积, 得到

$$p(\mu | \bar{y}, s^2) \propto \left[1 + \frac{n(\bar{y} - \mu)^2}{(n-1)s^2} \right]^{-n/2},$$

即

$$\frac{\mu - \bar{y}}{s/\sqrt{n}} \Big| \bar{y}, s^2 \sim t(n-1). \quad (11-4.4)$$

11.4.3 随机模拟方法

在大多数的实际问题中, 像上面正态分布那样能够得到感兴趣参数的边际后验分布是很少的. 然而我们可以通过随机模拟的方法获得边际后验分布的样本. 由公式(11-4.2), 得到 θ_1 的边际后验样本的抽样方法为:

第一步: 从 $p(\theta_2 | y)$ 中抽取 θ_2 ;

第二步: 从 $p(\theta_1 | \theta_2, y)$ 中抽取 θ_1 .

上述二步不断重复即可得到所需要的后验样本.

例 11.4.1 除了从 t 分布(11-4.4)直接抽取样本外, 我们也可按下面的二步简接获得 μ 的后验样本:

第一步: 按倒伽玛分布从 $p(\sigma^2 | y)$ 中抽取 σ^2 ;

第二步: 从 $p(\mu | \sigma^2, y) \sim N(\bar{y}, \sigma^2/n)$ 中抽取 μ .

有时可能遇到边际后验 $p(\theta_2|y)$ 无法得到显式表示, 特别是在多参数贝叶斯分析中, 这时经常采用Gibbs抽样法. 在两个参数 θ_1, θ_2 场合, 只需要改变一下上面的第一步. 整个算法变成

第一步: 给定 θ_1 的初始值;

第二步: 从 $p(\theta_2|\theta_1, y)$ 中抽取 θ_2 ;

第三步: 从 $p(\theta_1|\theta_2, y)$ 中抽取 θ_1 .

将此抽取过程重复进行, 即得到一系列基于后验分布的 θ_1 与 θ_2 的后验样本. 为保证独立性, 在使用之前应舍去没有达到平衡状态的那些样本.

最后, 若上述一维的边际后验分布或条件后验分布不易抽样, 则可以采用近似的离散化格式点抽样方法, 它也适用于二维分布的抽样, 其实施方法见下面的例子.

11.4.4 一个实例

除正态分布等少数模型之外, 一般多参数模型都无法得到后验分布的显式表示. 在实际应用中, 经常采用随机模拟的方法来解决这类问题. 下面给出一个在新药开发中动物试验的实例.

例 11.4.2 在药物以及其他一些化学合成剂的开发过程中, 需做毒性测试, 即在一批动物身上注射不同剂量的药物, 设动物的反应由两个对立的结果来描述, 例如“生存”或“死亡”. 此类试验的数据可以表示为

$$(x_i, n_i, y_i), i = 1, 2, \dots, k,$$

其中 k 为动物的分组数, n_i 表示第 i 批动物的动物个数, x_i 表示第 i 批动物接受的剂量水平(通常以对数形式出现), y_i 表示第 i 批动物服用剂量 x_i 后出现阳性反应的动物数(如“死亡”或“有肿块”的动物数). 现有一批动物共20只分为4组, 每组注射相同的剂量. 具体数据见表11.1. 如何根据试验数据判断该药物的毒性呢?

解 我们在无信息先验下分步讨论模型的建立与贝叶斯分析.

模型建立

对于第 i 批动物, n_i 个动物样本的试验结果可认为相互独立, 由此可导出二

表 11.1 动物的注射药物后的阳性反应

剂量(log(g/ml))	动物个数	死亡个数
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

项分布抽样模型

$$y_i | \theta_i \sim \text{Bin}(n_i, \theta_i),$$

其中 θ_i 死亡率. 同时 $\theta_1, \dots, \theta_4$ 也可认为相互独立, 且在许多场合中可假设 θ_i 与 x_i 有如下线性关系

$$\text{logit}(\theta_i) = \log\left(\frac{\theta_i}{1 - \theta_i}\right) = \alpha + \beta x_i. \quad (11-4.5)$$

回归分析

为考查动物的死亡率与接受药物新剂量的关系(11-4.5)是否合理, 我们考查 $\text{logit}(y_i/n_i)$ 对 x_i , $i = 1, 2, 3, 4$ 的回归关系. 由于 $y_1 = 0$ 和 $y_4 = 5$ 时无法求得 $\text{logit}(y_i/n_i)$, 故适当微调: $y_1 = 0.01$, $y_4 = 4.99$. R程序如下

```
> logit<-function(x){
  y=log(x/(1-x))
  return(y)
}
> bioassay<-data.frame(
> x <- c(-0.86, -0.30, -0.05, 0.73),
> n <- c(5, 5, 5, 5),
> y <- c(0.01, 1, 3, 4.99),
> r <- logit(y/n))
> plot(x,r)
> lm.bioassay<-lm(formula = r~x)
> abline(lm.bioassay)
```

```
> summary(lm.bioassay)
```

回归分析的结果为:

```
Call:
lm(formula = r ~ x)
Residuals:
    1      2      3      4 
-0.2190  0.2572  0.1069 -0.1450 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6870     0.1383   4.967 0.038226 *
x             7.7681     0.2368  32.810 0.000928 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2707 on 2 degrees of freedom
Multiple R-Squared:  0.9981,    Adjusted R-squared:  0.9972 
F-statistic: 1076 on 1 and 2 DF,  p-value: 0.0009277
```

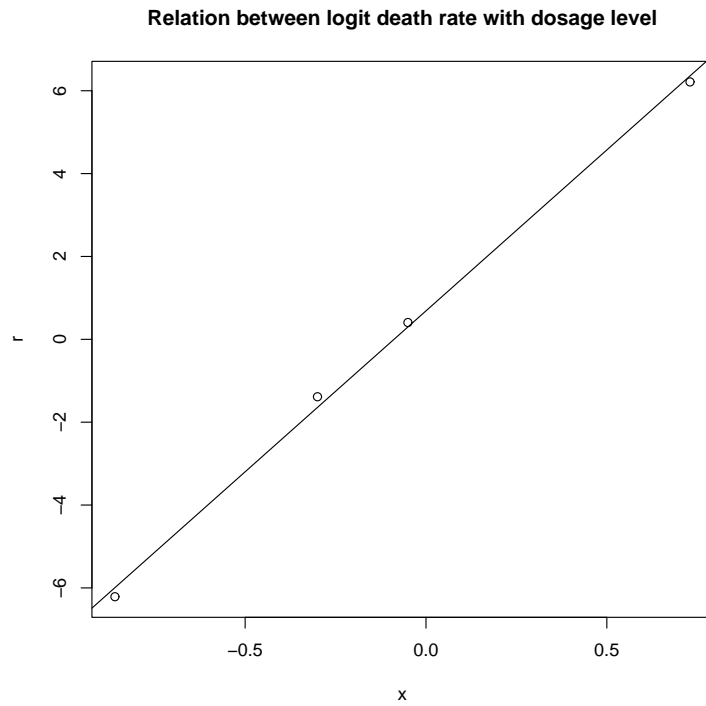
图11.5和回归分析的结果都表明上述假设是合理的。且得到 α 和 β 的估计分别为 $\hat{\alpha} = 0.69$ 和 $\hat{\beta} = 7.77$, 标准误差分别为0.1和0.24。

贝叶斯估计

若关于参数 α 和 β 没有可以利用的先验信息, 则采用无信息先验 $p(\alpha, \beta) \propto 1$ 。这时后验分布即为似然函数

$$\begin{aligned} p(\alpha, \beta | y, n, x) &\propto p(\alpha, \beta) p(y | \alpha, \beta) \\ &\propto \prod_{i=1}^k \left(\frac{e^{\alpha + \beta x_i}}{1 + e^{\alpha + \beta x_i}} \right)^{y_i} \left(\frac{1}{1 + e^{\alpha + \beta x_i}} \right)^{n_i - y_i} \quad (11-4.6) \end{aligned}$$

我们用后验众数作为参数 α 和 β 的点估计, 也即极大似然估计, 这可直接利用软件包stats4中的函数mle()求得, 下面的R程序先定义后验密度函数和负对数似然函数, 最后调用函数mle(), 具体代码如下:

图 11.5 由 θ 的后验分布产生的1000个随机样本的直方图.

```

> bioassay.post<-function(alpha=0.1,beta=5){
> k<-4
> x <- c(-0.86, -0.30, -0.05, 0.73)
> n <- c(5, 5, 5, 5)
> y <- c(0, 1, 3, 5)
> prod<-1
> prod <- prod((exp(alpha+beta*x)/(1+exp(alpha+beta*x[i]))))^y
>          *(1/(1+exp(alpha+beta*x)))^(n-y))
> return(prod)}
> mlpost<-function(alpha=0.1,beta=5){-log(bioassay.post(alpha,beta))}
> mle(mlpost)

```

得到的贝叶斯估计(极大似然估计)为 $\hat{\alpha} = 0.85$ 和 $\hat{\beta} = 7.75$.

后验密度及离散化抽样

由(11.4.4)无法得到 α 和 β 的后验分布, 因而无法得到相应的后验样本. 我们这里介绍一种连续密度的格子点离散化方法. 我们先画出在取值范围 $(\alpha, \beta) \in [-5, 10] \times [-10, 40]$ 的等高线图, 等高线分为0.05, 0.15, \dots , 0.95共10水平, **R**程序如下:

```
> modedensity<-bioassay.post(0.87,7.91)
> alphax<-seq(-5,10,length=1000)
> betay<-seq(-10,40,length=1000)
> post<-outer(alphax,betay,'bioassay.post')
> par(mfrow=c(1,2))
> contour(alphax,betay,post,
           levels=seq(0.05,0.95,length=10)
           *modedensity, xlim=c(-5,10),
           ylim=c(-10,40), xlab=quote(alpha),
           ylab=quote(beta), drawlabels= FALSE)
```

运行得到图11.8(a), 此图表明 α 与 β 有一定的正相关性.

对联合后验密度在 $(\alpha, \beta) \in [-5, 10] \times [-10, 40]$ 的 1000×1000 个格子点处离散化并抽样的步骤为:

- 1) 计算 $p(\alpha, \beta|y, n, x)$ 在所有格子点处的值;
- 2) 正则化使 $\sum_{\alpha} \sum_{\beta} p(\alpha, \beta|y) = 1$ 得到离散的联合后验分布列.
- 3) 由公式 $p(\alpha|y) = \sum_{\beta} p(\alpha, \beta|y)$ 得到 α 的边际后验分布;
- 4) 从离散的 $p(\alpha|y)$ 中抽取1000个 α ;
- 5) 对抽得的每一个 α , 由离散化条件分布 $p(\beta|\alpha, y)$ (也需要正则化)抽取相应的 β 值.

获得1000个后验样本点及相应的散点图(见图11.8)的**R**程序如下(续上):

```
> post<-post/sum(post)
> posta<-apply(post,MARGIN=1,FUN=sum)
```



```

> w<-posta/sum(posta)
> n<-1000
> ra<-rep(0,n)
> rb<-rep(0,n)
> for(j in 1:n){
  ra[j]<-sample(alphax,1,replace=T,prob=w)
  postb<-bioassay.post(ra[j],betay)
  wb<-postb/sum(postb)
  rb[j]<-sample(betay,1,replace=T,prob=w)
}
> plot(ra, rb, xlim=c(-5,10), ylim=c(-10,40),
  xlab=quote(alpha), ylab=quote(beta))

```

存活率为50%的剂量的估计

在此类生物鉴定试验中, 人们通常对导致50%存活率的剂量大小感兴趣, 记为 $LD50$. 在本例的logit模型中, 由

$$E\left(\frac{y_i}{n_i}\right) = \text{logit}^{-1}(\alpha + \beta x_i) = 0.5$$

解出 x_i 即得 $LD50 = -\alpha/\beta$. 利用上面抽取的 α, β 的1000个后验样本, 可以得到 $LD50$ 的(离散)后验分布, 其频率直方图图(见11.7)可由下面的R程序得到:

```

> ld50 <- -ra/rb
> hist(ld50,freq=FALSE,breaks=1000,xlim=c(-0.8,0.5),
  axes = TRUE, xlab="LD50",main="")

```

α, β 及 $LD50$ 的后验分位数如表11.2. 由这些结果可以得到动物的死亡率与接受的剂量成正比(因为 $\beta > 0$), 而有50%受试动物死亡的剂量为 $\exp(-0.11) = 0.90(\text{g/ml})$.

■

§11.5 分层贝叶斯统计分析

许多实际问题都会涉及多个参数, 而且这些参数会呈现出某种相关性. 统

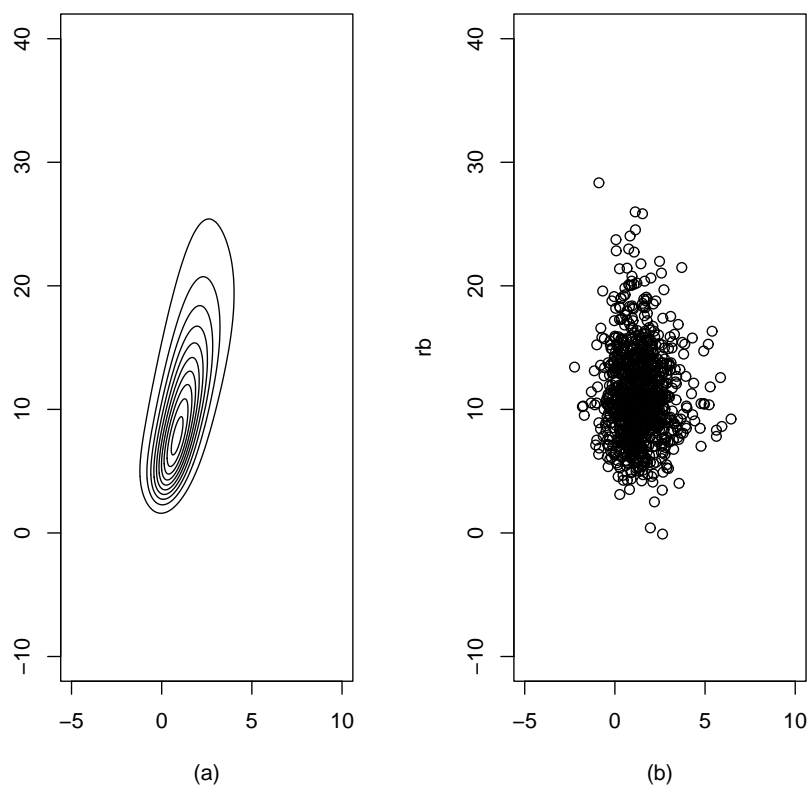


图 11.6 (a) α, β 基于后验密度的等高图 (b) 从后验密度中重新抽取 1000 对样本的散点图.

计上可以用一个联合概率分布来刻画参数之间的相依性. 例如, 在一个心脏病治疗的研究中, 考查 J 个医院使用某种药物后的存活率 $\theta_j, j = 1, 2, \dots, J$. 可以认为由数据得到的这些 θ_j 的估计应该是相互联系的. 在贝叶斯统计分析中, 这种参数间的相关性可以通过假设 $\theta_j, j = 1, 2, \dots, J$ 为来自一个共同的先验分布 (称为参数的总体分布) 的样本来实现, 即 $\theta | \phi \stackrel{iid}{\sim} p(\theta | \phi)$, 其中 ϕ 为未知超参数, 其本身有先验分布 $p(\phi)$. 这就是分层贝叶斯建模的思想.

这一节主要以数据 $y_j | \theta_j$ 为正态分布 $N(\theta_j, \sigma_j^2)$, 其中方差 σ_j^2 已知, 均值参数具有正态共轭先验 $\theta_j | \phi \sim N(\mu, \tau^2)$ 为例介绍分层模型的贝叶斯推断及其应用.

表 11.2 $\alpha, \beta, LD50$ 的后验分位数

	2.5%	25%	50%	75%	97.5%
α	-0.6	0.6	1.3	2.0	4.1
β	3.5	7.5	11.0	15.2	26.0
LD50	-0.28	-0.16	-0.11	-0.06	0.12

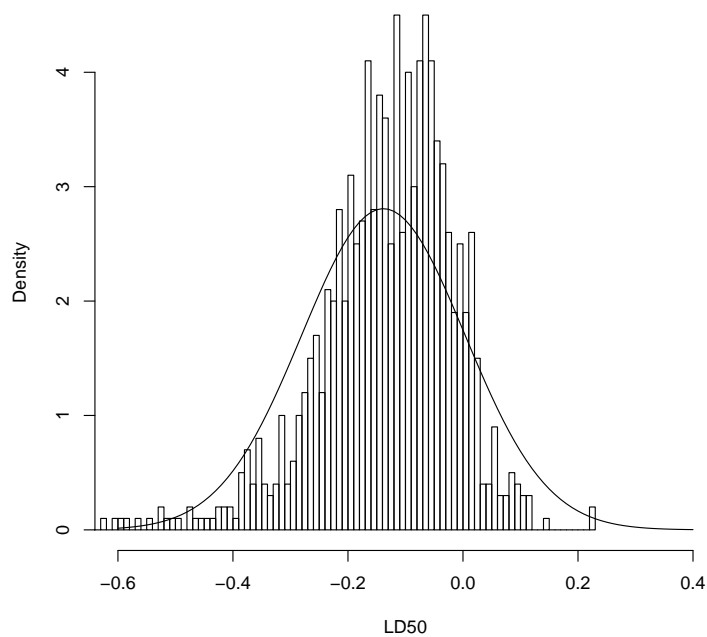


图 11.7 LD50 的后验频率直方图.

11.5.1 分层模型的建立及其贝叶斯推断

模型的建立

考虑 J 组试验, 由试验 j 得到数据 $(y_{j1}, \dots, y_{jn_j})$, 并综合为一个统计量 y_j (通常为充分统计量). 试验 j 所涉及的参数(向量)为 θ_j . 分层贝叶斯模型由三部分

组成:

1) 数据的分布(似然函数):

$$y_j | \theta_j \sim p(y | \theta_j)$$

令 $y = (y_1, y_2, \dots, y_J)$, $\theta = (\theta_1, \theta_2, \dots, \theta_J)$, 从而得到

$$y | \theta \sim \prod_{j=1}^J p(y_j | \theta_j). \quad (11-5.1)$$

2) 参数 θ_j 的先验分布: $\theta_j, j = 1, 2, \dots, J$ 为来自同一共分布 $p(\theta | \phi)$ 的样本, 因此

$$\theta | \phi \sim \prod_{j=1}^J p(\theta_j | \phi). \quad (11-5.2)$$

3) 超参数的先验分布:

$$\phi \sim p(\phi). \quad (11-5.3)$$

实际上这是一个双层贝叶斯模型, 我们还可引入更多层次的贝叶斯模型. 它也可视为一个多参数的贝叶斯模型, 但与我们前面讨论的多参数模型不同的是:

- 1) 在多参数的贝叶斯模型中超参数 ϕ 是通过历史数据估计(这时称为经验贝叶斯分析)或通过专家经验给定, 而在多层贝叶斯模型中 (θ, ϕ) 都是模型的参数, 尽管 θ 为主要关心的参数;
- 2) 在多参数的贝叶斯模型中参数 $\theta_1, \theta_2, \dots, \theta_J$ 通常假设为独立的, 没有相关的结构, 而在多层贝叶斯模型中 $\theta_1, \theta_2, \dots, \theta_J$ 之间存在着一种相关性, 这种相关性是通过其先验分布来刻划的.

由(11-5.1)、(11-5.2)和(11-5.3)得 (θ, ϕ) 的联合后验密度

$$\begin{aligned} p(\theta, \phi | y) &\propto p(\theta, \phi) p(y | \theta, \phi) \\ &\propto p(\phi) p(\theta | \phi) p(y | \theta). \end{aligned} \quad (11-5.4)$$

上式第二行是由于 y 仅依赖于 θ , 或者说 ϕ 仅通过 θ 影响 y .

θ 的统计推断

由多参数贝叶斯模型, 我们主要关心参数 θ 的统计推断和某一试验 j 下 y 的

预测. 由(11-5.4), θ 的后验分布为

$$p(\theta|y) \propto \int p(\theta|\phi, y)p(\phi|y)d\phi \quad (11-5.5)$$

其中 $p(\theta|\phi, y)$ 对于共轭先验分布容易得到, 为给定 ϕ 下 θ_j 的共轭后验分布的乘积, 而由条件概率密度计公式, $p(\phi|y)$ 可表示为

$$p(\phi|y) = \frac{p(\theta, \phi|y)}{p(\theta|\phi, y)}. \quad (11-5.6)$$

此式可避开积分计算: $p(\phi|y) = \int p(\theta, \phi|y)d\theta$. 因此关于 θ 的推断(抽样)可按下面的二步进行:

第1步: 由后验边际分布 $p(\phi|y)$ 推断(抽取) ϕ ;

第2步: 视 ϕ 为已知, 由条件后验分布 $p(\theta|\phi, y)$ 推断(抽取) θ .

预测

通常人们可能关心两类后验预测, 其一是基于现有的数据(试验), 这是一类常见的后验预测; 其二是基于新的试验, 当试验环境(协变量)不同时考试这种预测.

- 基于现有试验的预测: 现有试验的效应为 $\theta = (\theta_1, \dots, \theta_J)$, 这时抽样步骤为:
 - 1) 从 $p(\phi|y)$ 中抽取 ϕ ;
 - 2) 对于给定 $j \in (1, 2, \dots, J)$, 从 $p(\theta_j|\phi, y)$ 中抽取 θ_j ;
 - 3) 从 $p(y|\theta_j)$ 中抽取 \tilde{y} .
- 基于新试验的预测: 这时需要先获得试验的效应 $\tilde{\theta}$, 抽样步骤变为
 - 1) 从 $p(\phi|y)$ 中抽取 ϕ ;
 - 2) 从参数(效应)的总体分布 $p(\theta|\phi)$ 中抽取新的参数 $\tilde{\theta} = (\tilde{\theta}_1, \dots, \tilde{\theta}_J)$;
 - 3) 从 $p(y|\tilde{\theta})$ 中抽取 \tilde{y} .

11.5.2 N-N模型与应用

问题的叙述

数据与参数都服从正态分布的多层贝叶斯模型称为N-N模型. 我们先从经典的统计分析中引出这个问题.

考查 J 个试验, 测得数据为 y_{ij} , 设

$$y_{ij} \stackrel{iid}{\sim} N(\theta_j, \sigma^2), i = 1, 2, \dots, n_j (j = 1, 2, \dots, J) \quad (11-5.7)$$

其中方差 σ^2 已知. 则样本均值 $\bar{y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$ 为 θ_j 的充分统计量, 有分布

$$\bar{y}_{.j} | \theta_j \sim N(\theta_j, \sigma_j^2), \quad (11-5.8)$$

其中 $\sigma_j^2 = \frac{\sigma^2}{n_j}$.

现在我们考查某个特定 θ_j 的估计. 我们可想到两种估计:

- 1) 使用单个 $y_{.j}$ 进行估计: $\hat{\theta}_j = y_{.j}$ 这时当 n_j 很小时显然是不合理的, 因为它的精度会很低.
- 2) 使用合并数据的估计, 即将 J 个试验的条件和对像没有多少差异, 即认为 $\theta_1 = \dots = \theta_J$, 则

$$\hat{\theta}_j = \bar{y}_{..} = \frac{\sum_{j=1}^J \frac{1}{\sigma_j^2} \bar{y}_{.j}}{\sum_{j=1}^J \frac{1}{\sigma_j^2}}.$$

到底选用哪一个, 可通过 J 个组(试验)下 $\theta_j, j = 1, 2, \dots, J$ 的差的方差分析(F 检验)进行. 设 τ^2 为 $\theta_j, j = 1, 2, \dots, J$ 的(先验)方差. 为方便起见, 在此仅考虑 J 组试验是均衡的, 即 $n_j = n, \sigma_j^2 = \sigma^2, j = 1, 2, \dots, J$. 则理论上, F 检验的方差分析表可表示为

由此我们得出:

- 1) 如果组间平方和与组内平方和之比显著大于1, 就认为 $\theta_j, j = 1, 2, \dots, J$ 之间有显著差异, 这时就取 $\hat{\theta}_j = y_{.j}$;
- 2) 如果组间平方和与组内平方和之比并不显示大于1, 就认为 $\theta_j, j = 1, 2, \dots, J$ 之间没有有显著差异, 即 F 检验无法拒绝 $H_0: \tau = 0$, 这时就取 $\hat{\theta}_j = \bar{y}_{..}$.

表 11.3 F 检验理论上的方差分析表

	自由度	SS	MS	$E(MS \sigma^2, \tau)$
组间	$J - 1$	$\sum_i \sum_j (\bar{y}_{.j} - \bar{y}_{..})^2$	$\frac{SS}{J - 1}$	$n\tau^2 + \sigma^2$
组内	$J(n-1)$	$\sum_i \sum_j (\bar{y}_{ij} - \bar{y}_{.j})^2$	$\frac{SS}{J(n-1)}$	σ^2
总和	$Jn-1$	$\sum_i \sum_j (\bar{y}_{ij} - \bar{y}_{..})^2$	$\frac{SS}{Jn-1}$	σ^2

实际上, 我们还可提出第三种估计

$$\hat{\theta}_j = \lambda_j \bar{y}_{.j} + (1 - \lambda_j) \bar{y}_{..},$$

其中 $0 \leq \lambda_j \leq 1$. 它是前二种估计的加权平均, 它可视为 J 个组平均之间有一定的联系或存在一定的相依结构. 与方差分析表比较发现, $\lambda_j = 1$ 对应于 τ^2 很大, 在此视为 ∞ , $\lambda_j = 0$ 对应于 $\tau^2 = 0$, 而 $0 \leq \lambda_j \leq 1$ 对应一个较适中的 τ^2 . 采用分层贝叶斯模型恰能获得这种折衷的估计, 且能包含前二种特殊的情形, 由此进一步假设 $\theta_j, j = 1, 2, \dots, J$ 为来自正态分布 $N(\mu, \tau^2)$ 的样本, 而超参数 (μ, τ) 服从先验 $p(\mu, \tau^2)$. $\theta_j, j = 1, 2, \dots, J$ 之间的相依性就是通过引入这个共同的先验分布来实现的.

为方便起见令 $y_{.j} = y_j$, 则 N-N 的数据与参数分别为 $y = (y_1, y_2, \dots, y_J)$, $\theta = (\theta_1, \theta_2, \dots, \theta_J)$. 这样 N-N 分层贝叶斯模型可表示为

$$p(y|\theta) = \prod_{j=1}^J N(y_j|\theta_j, \sigma_j^2) \quad (11-5.9)$$

$$p(\theta|\mu, \tau) = \prod_{j=1}^J N(\theta_j|\mu, \tau^2) \quad (11-5.10)$$

$$(\mu, \tau) \sim p(\mu, \tau) \quad (11-5.11)$$

我们仅考虑 $p(\mu|\tau) \propto 1$, $p(\tau) \propto 1$, 所以 $p(\mu, \tau) = p(\mu|\tau)p(\tau) \propto 1$. 由此得

到 (θ, μ, τ) 的后验分布

$$\begin{aligned} p(\theta, \mu, \tau|y) &\propto \tau^{-J} \exp \left[-\frac{1}{2} \sum_j \frac{1}{\tau^2} (\theta_j - \mu)^2 \right] \\ &\quad \exp \left[-\frac{1}{2} \sum_j \frac{1}{\sigma_j^2} (y_j - \theta_j)^2 \right] \end{aligned} \quad (11-5.12)$$

下面我们列出一些结果

1) 给定 μ, τ, y 下, θ 的分布

$$\theta_j | \mu, \tau, y \sim N(\hat{\theta}_j, V_j), \quad (11-5.13)$$

其中

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2} y_j + \frac{1}{\tau^2} \mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}, \quad V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

2) 给定 y 下, μ, τ 的分布

$$\begin{aligned} p(\mu, \tau|y) &\propto \prod_{j=1}^J N(y_j | \mu, \sigma_j^2 + \tau^2) \\ &\propto \prod_{j=1}^J (\sigma_j^2 + \tau^2)^{-1/2} \exp \left(-\frac{(y_j - \mu)^2}{2(\sigma_j^2 + \tau^2)} \right) \end{aligned} \quad (11-5.14)$$

3) 给定 τ, y 下, μ 的分布

$$\mu | \tau, y \sim N(\hat{\mu}, V_\mu), \quad (11-5.15)$$

其中

$$\hat{\mu} = \frac{\sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau^2} y_j}{\sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau^2}}, \quad V_\mu = \frac{1}{\sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau^2}}$$

4) 给定 y 下, τ 的分布

$$\begin{aligned} p(\tau|y) &= \frac{p(\mu, \tau|y)}{p(\mu|\tau, y)} \\ &\propto \frac{\prod_{j=1}^J N(y_j | \hat{\mu}, \sigma_j^2 + \tau^2)}{N(\hat{\mu} | \hat{\mu}, V_\mu)} \end{aligned}$$

$$\propto V_{\mu}^{1/2} \prod_{j=1}^J (\sigma_j^2 + \tau^2)^{-1/2} \exp \left(-\frac{(y_j - \hat{\mu})^2}{2(\sigma_j^2 + \tau^2)} \right) \quad (11-5.16)$$

可见, θ_j 的估计恰为单个数据的估计 y_j 与 μ 的加权平均, 而 μ 的估计为各 $\theta_j, j = 1, 2, \dots, J$ 的“合并估计”. 若 $\tau = 0$, θ_j 的估计即为基于单个数据的估计, 而当 $\tau = \infty$ 时, θ_j 的估计即为 μ 的估计.

由此可得N-N模型的抽样方法:

- 1) 按格子点离散化方法由公式(11-5.16)从 $p(\tau|y)$ 抽取 τ ;
- 2) 按正态分布由公式(11-5.15)从 $p(\mu|\tau, y)$ 抽取 μ ;
- 3) 按独立正态性由公式(11-5.13)从 $N(\hat{\theta}_j, V_j)$ 抽取 $\theta_j, j = 1, 2, \dots, J$.

案例分析

例 11.5.1 SAT考试旨在真实地考察学生经过多年教育之后所获取的知识与能力, 同时极力避免因短期突击而带来的成绩提高. 为研究短期考前培训是否能提高SAT(学校智能测试)的成绩, 现对8所进行过此类培训的高中进行独立随机试验, 经过协方差调整(以消除其它因素的影响)后得到的数据如表11.4所示. 由于每个学校参加测试的学生数都至少有32人, 因此可以认为 y_j 具有正态近似, 并用样本方差作为 σ_j^2 的值. 现要研究8所学校短期考前培训的真实效果, 并进行比较.

解 我们用N-N模型逐步展开讨论:

- 1) 二个极端的估计: 考虑学校A的培训效应, 若认为8所学校没有关系, 则用单个的数据估计, 即 $\hat{\theta}_1 = 28$ (标准差为15); 若认为8个学校的培训效应没有差异, 则使用合并估计(pooled estimate):

$$\hat{\theta}_1 = \hat{\mu} = \frac{\sum_{j=1}^8 (y_j / \sigma_j^2)}{\sum_{j=1}^8 (1 / \sigma_j^2)} = 7.9 \text{ (标准差为 } 4.2 \text{)}.$$

- 2) 初步分析: 一些学校的培训呈现了一定的效果(18到28之间), 一些学校则效果较小, 还有的有相反的效果. 而且较大的标准误差意味着各平均效应 θ_j 的置信区间会有较大的重叠, 即统计上很难区分它们. 然而, 经典的统计分析却拒绝各 $\theta_j, j = 1, 2, \dots, J$ 相等的假设. 因此, 上面的二个估计

表 11.4 SAT成绩

学校	培训效果 估计值	培训效果估 计值标准误
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

都是不合理的. 下面用N-N多层贝叶斯模型给出一个介于二者之间的一个折衷的估计.

3) N-N多层贝叶斯分析: 套用N-N模型对此问题进行分析. 现在的模型为

$$y_j | \theta_j, \sigma_j^2 \sim N(\mu_j, \sigma_j^2)$$

其中 $\theta_j (j = 1, 2, \dots, 8)$ 即为8所学校各自短期培训的“真实”效果, 且

$$\theta_j | \mu, \tau \sim N(\mu, \tau^2),$$

其中未知超参数 μ 与 τ 相互独立, 并假设 $p(\mu, \tau) \propto 1$.

- τ 的后验分布: 利用网格法画出 τ 的边际后验密度 $p(\tau|y)$ 的函数, **R**程序如下

```
> y<-c(28, 8, -3, 7, -1, 1, 18 ,12)
> sd<-c(15, 10, 16, 11, 9, 11, 10, 18)
> v<-sd*sd
> tau<-c(0:3000)/100
> tausq<-tau*tau
> ptau.y<-rep(0,3001)
```

```

> vmu<-rep(0,3001)
> muhat<-rep(0,3001)
> for(i in (1:3001)){
  vmu[i]<-1/sum(1/(tausq[i]+v))
  muhat[i]<-vmu[i]*sum(y/(tausq[i]+ v))
  ptau.y[i]<-sqrt(vmu[i]*prod(1/(tausq[i]+v)))
    *prod(exp(-0.5*(y-muhat[i])
      *(y-muhat[i])/(tausq[i]+v)))
}
> plot(tau,ptau.y,type="l",yaxt="n",xlab=quote(tau))

```

得到图11.8. 由图11.8可知, τ 值趋近于0时最为合理, 且有

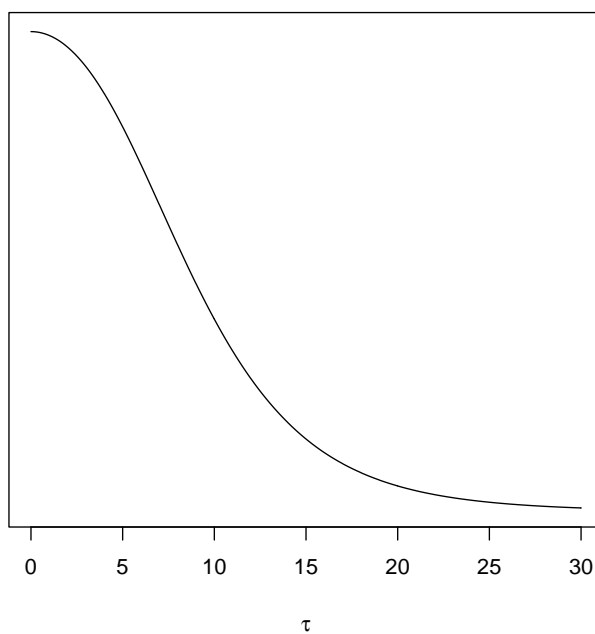


图 11.8 边际后验密度函数 $p(\tau|y)$.

$$Pr(\tau > 10) < 0.5, Pr(\tau > 25) \approx 0.$$

- 给定 τ 下各效应的平均水平与波动: 为进一步了解 τ 的性质, 现考虑在 τ 给定下的后验均值 $E(\theta_j|\tau, y)$ 及其相应标准差 $Sd(\theta_j|\tau, y)$. 经计算得到

$$E(\theta_j|\tau, y_j) = \frac{\frac{1}{\sigma_j^2}y_j + \frac{1}{\tau^2}\hat{\mu}}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}},$$

$$Sd(\theta_j|\tau, y_j) = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} + \left(\frac{\frac{1}{\tau^2}}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}} \right)^2 V_\mu,$$

其中

$$\hat{\mu} = \frac{\sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau^2} y_j}{\sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau^2}}, V_\mu^{-1} = \sum_{j=1}^J \frac{1}{\sigma_j^2 + \tau^2}.$$

根据上述公式对两者作图比较. **R**程序如下(接上段程序):

```
> eth.tauy<-matrix(rep(0,24008),8,3001)
> sdth.tauy<-matrix(rep(0,24008),8,3001)
> for (j in (1:8)){
>   for (i in (1:3001)){
eth.tauy[j,i]<-(y[j]/v[j]+muhat[i]/tausq[i])
/(1/v[j]+1/tausq[i])
sdth.tauy[j,i]<-sqrt(((1/tausq[i])/(1/v[j]+1/tausq[i]))
*((1/tausq[i])/(1/v[j]+ 1/tausq[i]))
*vmu[i]+1/(1/v[j]+1/tausq[i]))
}
}
> par(mfrow=c(1,2))
> taux<-matrix(rep(tau,8),8,byrow=T)
> matplot(t(taux),t(eth.tauy), ylim=(c(-5,30)),
type="l", xlab="tau",lty = 1:8, lwd = 1,col=1,
ylab="Estimate Treatment Effects",
main="Conditional posterior mean")
> School<-c("A","B","C","D","E","F","G","H")
> text(x=rep(20,8),y=t(eth.tauy)[2400,],School)
> matplot(t(taux),t(sdth.tauy), ylim=(c(0,20)),
```

```

type="l", xlab="tau", lty = 1:8, lwd = 1, col=1,
ylab="Posterior Standard Deviations",
main="Conditional posterior SD")
> text(x=rep(20,8), y=t(sdth.tauy)[2400,], School)

```

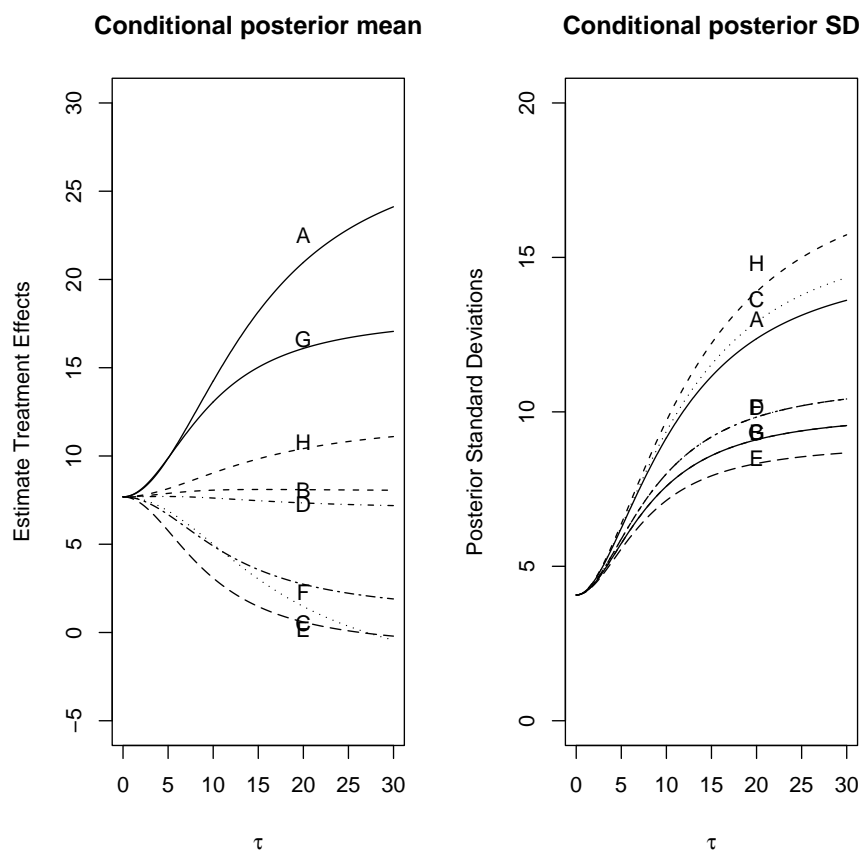


图 11.9 条件后验均值 $E(\theta_j|\tau, y_j)$ 与标准差 $Sd(\theta_j|\tau, y_j)$..

运行得到图11.9. 图11.9表明, 当 τ 取最合理的值(即为0)时, 8个不同学校的效应值 θ_j 的均值与标准差都几乎相同, 而随着 τ 不断变大, 它们之间的差异也变得明显起来, 且与各自最初的试验数据相接近. 因此, 仅根据 τ 无法得到满意的结果.

- 后验抽样: 下面的 \mathbf{R} 矩阵中 x 放置抽样的结果, 其中第1列 $x[, 1]$ 放置 $p(\tau|y)$ 的样本, 第2列 $x[, 2]$ 放置 $p(\mu|\tau, y)$ 的样本, 第3-10列 $x[, j]$, $j =$

3, 4, ..., 10 放置 $p(\theta_j | \tau, \mu, y), j = 1, 2, \dots, 8$ 的样本.

```
> m<-200
> ptau.y<-ptau.y/(sum(ptau.y))
> tausamp<-sample(tau,m,replace=T,prob=ptau.y)
> tausamp<-sort(tausamp)
> tauid<-tausamp*100 + 1
> x<-matrix(rnorm(m*10,0,1),m,10)
> x[,1]<-tausamp
> x[,2]<-muhat[tauid] + sqrt(vmu[tauid])*x[,2]
> for(j in (1:8)) {
  thmean<-(y[j]*x[,1]*x[,1]+v[j]*x[,2])/(v[j]+x[,1]*x[,1])
  thsd<-sqrt(v[j]*x[,1]*x[,1]/(v[j]+x[,1]*x[,1]))
  x[,j+2]<-thmean + thsd*x[,j+2]
}
> par(mfrow=c(1,2))
> hist(x[,2],breaks=c(-40:50),xlab=quote(mu),
      yaxt="n",main="")
> hist(x[,3],breaks=c(-20:60),xlab="Effect in School A",
      yaxt="n",main="")
```

画出了 μ 与 θ_1 的后验密度的直方图, 见图11.10.

- 后验推断: 对于每个 θ_j 的200个样本运用函数 `sort()` 进行排序后, 可以得到相应的五个分位数. 由表11.5不难发现, 根据200个样本, 8个学校实际培训效果的95%置信区间有很高的重叠性, 且其均值都处于5至10的范围之内.

■

§11.6 贝叶斯线性回归分析

11.6.1 模型的表示

本节主要就正态线性回归模型进行简单的讨论: 设 y 为响应变量, x_1, x_2, \dots, x_k 为 k 个预测变量, $\beta_1, \beta_2, \dots, \beta_k$ 为对应的回归系数. 对 n 个个体进行观察, 第 i 个响应变量与预测变量的值分别为 y_i 和 $x_i = (x_{i1}, x_{i2}, \dots, x_{ik})$,

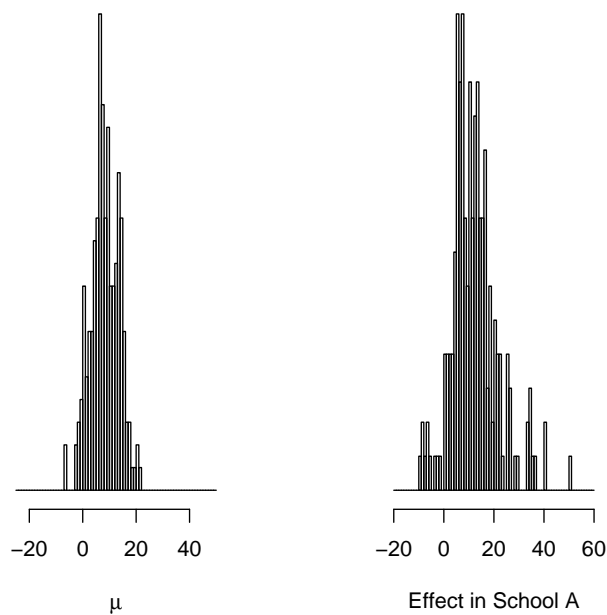


图 11.10 分别从 $p(\mu|\tau, y)$ 及 $p(\theta_1|\mu, \tau, y)$ 中抽取 200 个样本的频数直方图.

$i = 1, 2, \dots, n$. 记 $y = (y_1, y_2, \dots, y_n)'$, $\beta = (\beta_1, \beta_2, \dots, \beta_k)'$,

$$\mathbb{X} = (x_1, x_2, \dots, x_n)' = \begin{pmatrix} x_{11} & x_{12} & \dots & y_{1k} \\ x_{21} & x_{22} & \dots & y_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & y_{nk} \end{pmatrix},$$

这时正态回归模型可表示为

$$y|\beta, \sigma^2, \mathbb{X} \sim N_n(\mathbb{X}\beta, \sigma^2 \mathbf{I}_n), \quad (11-6.1)$$

或表示为

$$\begin{aligned} y_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \epsilon_i, i = 1, 2, \dots, n, \\ \epsilon &= (\epsilon_1, \dots, \epsilon_n) \stackrel{iid}{\sim} N(0, \sigma^2). \end{aligned} \quad (11-6.2)$$

表 11.5 θ_j 的后验分位数

学校	2.5%	25%	median	75%	97.5%
A	-2.	6	9	15	32
B	-5	4.	8	11	21
C	-9	2	6	10	19
D	-5	4	7	11	20
E	-9	1	5	8	15
F	-11	2	6	10	18
G	-1	6	9	13	24
H	-7	4	8	12	30

其中 y 称为观测向量, \mathbb{X} 称为设计阵, $\theta = (\beta, \sigma^2)$ 均为未知参数向量, \mathbf{I}_n 为单位矩阵, $N_n(\mu, A)$ 为 n 元正态分布, μ 为其均值向量, A 为其协方差矩阵.

11.6.2 后验分布

为了进行贝叶斯分析, 还需要给出 $\theta = (\beta, \sigma^2)$ 的先验分布, 在此我们仅对未知参数进行无信息先验假设, 即

$$p(\beta, \sigma^2) \propto (\sigma^2)^{-1}. \quad (11-6.3)$$

基于多元正态分布的性质, 此模型的贝叶斯推断可以借鉴多参数模型的情况. 现将后验分布表示为

$$p(\beta, \sigma^2 | y) = p(\sigma^2 | y) p(\beta | \sigma^2, y). \quad (11-6.4)$$

其中

$$\beta | \sigma^2, y \sim N_n(\hat{\beta}, V_\beta \sigma^2), \quad (11-6.5)$$

$$\sigma^2 | y \sim \text{IGa}((n - k)/2, S/2), \quad (11-6.6)$$

$\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$, $V_\beta = (\mathbb{X}'\mathbb{X})^{-1}$, $S = (y - \mathbb{X}\hat{\beta})^T(y - \mathbb{X}\hat{\beta})$. 不难发现, $\hat{\beta}$ 正是 β 的最小二乘估计.

因此按随机模拟的方法, 先由 $p(\sigma^2|Y)$ 抽取 σ^2 , 再从 $p(\beta|\sigma^2, y)$ 中抽取 β , 便能得到未知参数以及其函数的后验模拟值. 这可根据多元正态分布和逆伽玛分布自行编程获得 (β, σ^2) 的后验样本. 但是在无信息先验假设下我们也可直接利用R中的**LearnBayes**程序包(可在R社区下载)内的函数函数**blinreg()**完成 (β, σ^2) 的后验抽样.

11.6.3 回归拟合

有了 (β, σ^2) 的后验样本, 就可得到 β 的贝叶斯估计(如后验样本众数或后验样本均值), 记为 $\hat{\beta}$, 由此可得给定预测变量 x^* 处响应变量 y 的值

$$\hat{y} = x^* \hat{\beta}.$$

如果 β^* 为 β 的后验抽样, 则 $x^* \beta^*$ 就是 $x^* \beta$ 的边际后验的抽样. **LearnBayes**程序包中的函数**blinregexpected()**可以用于获得这样的样本.

11.6.4 后验预测

由多参数贝叶斯模型知, 给定预测变量 x^* 处, y 的后验预测分布为

$$p(\tilde{y}|y) = \int p(\tilde{y}|\beta, \sigma^2) p(\beta, \sigma^2|y) d\beta d\sigma^2. \quad (11-6.7)$$

因此 y 的后验预测样本可在上面获得的 (β, σ^2) 的后验样本的基础上, 再从正态分布 $N(x^* \beta, \sigma^2)$ 抽取 \tilde{y} 得到. **LearnBayes**程序包中的函数**blinregpred()**可以用于获得这样的后验预测样本. 显然, x^* 处, y 的回归拟合均值与预测均值都为 $x^* \beta$.

下面通过一个实例来说明这些函数的使用.

例 11.6.1 (Ramsey and Schafer, 1997)**LearnBayes**程序包中的**birdextinct**数据集为过去几十年中在英国周围的16个岛屿上收集的62种鸟的四类数据:

- 在岛上的平均灭绝时间(TIME);
- 平均筑巢数(NESTING);
- 种群规模(SIZE), 分为“大”(用1表示)与“小”(用0表示)两类;

- 栖息状态(STATUS), 分为“迁徙”(用1表示)与“久居”(用0表示)两类.

由命令

```
> data(birdextinct)
> attach(birdextinct)
> birdextinct
```

得到数据(仅前后一部分)

	species	time	nesting	size	status
1	Sparrowhawk	3.030	1.000	0	1
2	Buzzard	5.464	2.000	0	1
3	Kestrel	4.098	1.210	0	1
4	Peregrine	1.681	1.125	0	1
...
60	Starling	41.667	11.620	1	1
61	Pied_flycatcher	1.000	1.000	1	0
62	Siskin	1.000	1.000	1	1

研究目的是找出该地区鸟类的灭种时间与其余三个量之间的关系.

解

- 1) 预测变量的显著性: 按习惯, 用 y 表示响应变量TIME, x_1 表示预测变量NESTING, x_2 表示预测变量SIZE, x_3 表示预测变量STATUS. 由于前期分析中发现变量 y 严重右偏, 因此对其进行对数处理. 最终将此问题归为线形回归模型

$$\begin{aligned}\log(y_i) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \\ \epsilon_i &\stackrel{iid}{\sim} N(0, \sigma^2)\end{aligned}$$

首先用函数lm()进行最小二乘拟合. R命令为

```
> logtime=log(time)
> fit=lm(logtime ~ nesting+size+status,
          data=birdextinct, x=TRUE, y=TRUE)
> summary(fit)
```

其中 $x=TRUE$, $y=TRUE$ 是为了让设计矩阵和响应变量成为`fit`这个结构的一部分, 便于在后面的函数中引用. 输出的主要结果为

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.43087     0.20706   2.081 0.041870 *
nesting      0.26501     0.03679   7.203 1.33e-09 ***
size        -0.65220     0.16667  -3.913 0.000242 ***
status       0.50417     0.18263   2.761 0.007712 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6524 on 58 degrees of freedom
Multiple R-Squared:  0.5982,    Adjusted R-squared:  0.5775
F-statistic: 28.79 on 3 and 58 DF,  p-value: 1.577e-11
```

结论: 回归方程为

$$y = 0.43087 + 0.2650x_1 - 0.6522x_2 + 0.5042x_3.$$

且筑巢数 $NESTING(x_1)$ 是高度显著的, 表明筑巢数越多, 种类灭绝的时间越长; 种群规模 $SIZE(x_2)$ 和栖息状态 $STATUS(x_3)$ 也显著, 但稍差一点, 表明大的鸟类其灭绝的时间短; 而迁徙的鸟类其灭绝的时间长.

2) 产生 $\theta = (\beta, \sigma)$ 的后验样本: 命令

```
> theta.sample <- blinreg(fit$y, fit$x, 5000)
```

得到 β, σ 的5000个后验样本.

说明: 函数`blinreg(y, X, m)`所需输入的变量为: 观测向量 y , 结构矩阵 X 以及样本量 m . 此函数的返回值分为两部分: 第一部分为 β 的 $m \times k$ 矩阵样本, 其每一行分别代表该次抽样的 β_i 值($i = 0, 1, 2, \dots, k$), 第二部分则为 m 个 σ 的样本值, 且这两部分的值被赋予变量名`beta`和`sigma`. 在此 $m = 5000, k = 3$.

命令

```
> par(mfrow=c(2,2))
```

```

> hist(theta.sample$beta[,2], main="NESTING",
      xlab=expression(beta[1]))
> hist(theta.sample$beta[,3], main="SIZE",
      xlab=expression(beta[2]))
> hist(theta.sample$beta[,4], main="STATUS",
      xlab=expression(beta[3]))
> hist(theta.sample$sigma, main="ERROR SD",
      xlab=expression(sigma))

```

得到 $\beta_1, \beta_2, \beta_3$ 和 σ 的直方图(见图11.11).

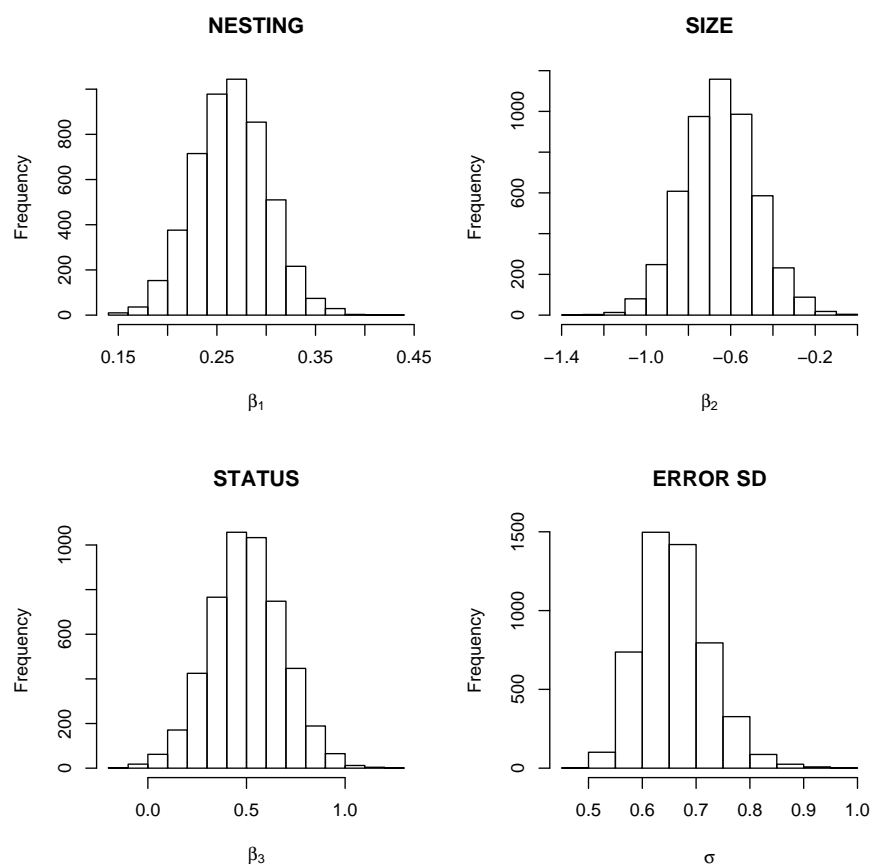


图 11.11 $\beta_1, \beta_2, \beta_3$ 和 σ 的后验频数直方图.

- 3) 对参数的概括: 根据我们需要我们可以用后验样本对未知参数作出推断, 例如使用函数`apply()`及`quantile()`计算 β 与 σ 的后验样本的5%、50%、95%分位数. **R**命令与结果如下

```
> apply(theta.sample$beta, 2, quantile, c(.05, .5, .95))
      X(Intercept) Xnesting      Xsize      Xstatus
5%      0.0885280 0.2049362 -0.9288856 0.1987654
50%      0.4275055 0.2642518 -0.6503313 0.4998742
95%      0.7789398 0.3247912 -0.3739433 0.8085352

> quantile(theta.sample$sigma, c(.05, .5, .95))
      5%      50%      95%
0.5676048 0.6545287 0.7729843
```

若用`summery`命令观察最小二乘估计的结果不难发现, 各未知参数的后验中位数与该结果基本一致, 其原因在于, 本例的贝叶斯推断采用了无信息先验假设.

另外我们还可得到已知预测变量 $x_i, i = 1, 2, \dots, k$ (本例 $k = 3$)时回归均值与预测值, 这部分作为练习(习题11.8)请读者完成. ■

第十一章习题

11.1 试证明: 对于二项分布 $\text{Bin}(n, \theta)$ 中的比率 θ , 基于杰弗莱原则的先验分布为贝塔分布 $\text{Beta}(1/2, 1/2)$.

11.2 考虑女性的出生比率 θ , 设 θ 仅有二种可能: $\theta = 0.5$ 与 $\theta = 0.485$, 且 θ 的先验分布是等可能的. 令 y 为 n 个出生的新生儿中女性的数目, 试分别在1) $n = 100, y = 48$, 2) $n = 1000, y = 480$ 两种场合, 分别求 θ 的后验分布、后验均值(贝叶斯估计)和后验方差. 并对两者的差异进行说明.

11.3 设 $y|\theta \sim \text{Bin}(n, \theta)$, θ 先信息先验分布 $\text{Beta}(1, 1)$, 试根据已有的试验结果 n, y 预测下一次试验成功($\tilde{y} = 1$)与失败($\tilde{y} = 0$)的概率.

11.4 设 $y|\theta \sim \text{Bin}(n, \theta)$, θ 先信息先验分布 $\text{Beta}(1, 1)$, 考查4种观测数据: $n = 5, y = 3$ 、 $n = 20, y = 12$ 、 $n = 100, y = 60$ 、 $n = 1000, y = 600$,

- 1) 求4种场合参数 θ 的经典极大似然估计;
- 2) 用R编程计算4种场合参数 θ 的贝叶斯估计和精度, 并画图予以说明.

11.5 在例11.3.4中, 若取 θ 的后验分布为共轭分布 $\text{Beta}(50, 25)$, 求 θ 、 $\phi = \frac{1-\theta}{\theta}$ 和 $\text{logit}(\theta) = \log(\theta/(1-\theta))$ 后验均值与95%后验置信区间.

11.6 给定如下的贝叶斯模型: 对于 $j = 1, 2$

$$\begin{aligned} y_{j1}, \dots, y_{jn_j} | \mu_j, \sigma_j^2 &\sim N(\mu_j, \sigma_j^2), \\ p(\mu_j, \sigma_j^2) &\propto \sigma_j^{-2}, \end{aligned}$$

且 (μ_1, σ_1^2) 与 (μ_2, σ_2^2) 独立. 证明 $(s_1^2/s_2^2)/(\sigma_1^2/\sigma_2^2)$ 的后验分布为 F 分布:

$$\left. \frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2} \right| \bar{y}_j, s_j^2, j = 1, 2 \sim F(n_1 - 1, n_2 - 1).$$

11.7 设 $y = (y_1, \dots, y_n) \stackrel{iid}{\sim} N(\mu, \sigma^2)$, 其中 μ 和 σ^2 均未知, 其先验取为无信息先验 $p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$.

- 1) 从正态分布 $N(100, 1)$ 中产生1000个随机数;
- 2) 求参数 μ 和 σ^2 的贝叶斯估计(后验众数);
- 3) 作出后验密度函数的等高线图;

- 4) 用例(11.4.1)的算法产生 (μ, σ^2) 的1000个后验样本, 并作出散点图;
- 5) 用后验密度函数 $p(\mu, \sigma^2|y)$ 的格子点离散近似方法产生 (μ, σ^2) 的1000个后验样本, 并作出散点图, 并与上一散点图并列放置在一张图上进行比较.

11.8 基于例11.6.1, 取如下4组预测变量(协变量)

表 11.6 协变量

编号	x_1	x_2	x_3
A	4	0	0
B	4	1	0
C	4	0	1
D	4	1	1

- 1) 对于4组协变量分别得到回归均值 $x^*\beta$ 的样本, 并在同一个图中画出它们的直方图(使用命令`blinregexpected()`);
- 2) 对于4组协变量分别得到预测响应变量的预测值 \hat{y} 的样本, 并在同一个图中画出它们的直方图(使用命令`blinregpred()`);

附录 A 秩与结的介绍

设有独立同分布的样本 X_1, X_2, \dots, X_n , 不妨假设总体是连续型随机变量, 从而以概率1保证样本单元 X_1, X_2, \dots, X_n 互不相等, 将样本单元有小到大排列成 $X_{(1)} < X_{(2)} < \dots < X_{(n)}$. 若 $X_i = X_{(R_i)}$, 则称 $X_i (i = 1, 2, \dots, n)$ 在 X_1, X_2, \dots, X_n 中的秩为 R_i , 简称 X_i 的秩为 R_i , $R_i = 1, 2, \dots, n$. 秩方法的基本思想是, 用 X_i 的秩 R_i 代替 X_i 作统计推断. $R = (R_1, R_2, \dots, R_n)$ 以及由 R 构造的任意的统计量都称为秩统计量.

R 服从离散分布, 它取 $n!$ 个值. 由于样本 X_1, X_2, \dots, X_n 独立同分布, 所以 R 取任意一组值 (r_1, r_2, \dots, r_n) 的概率是 $1/n!$, 其中 (r_1, r_2, \dots, r_n) 是 $(1, 2, \dots, n)$ 的任意一个排列, 这说明 R 服从均匀分布. 由此可见, 秩统计量的分布与总体服从什么样的分布无关, 这就是称秩方法为非参数方法的原因.

由于 R 服从均匀分布, 所以单个样本的秩 $R_i (i = 1, 2, \dots, n)$ 也服从均匀分布: $P(R_i = r) = \frac{1}{n}, i = 1, 2, \dots, n$, 从而有:

定理1 对任意的 $i = 1, 2, \dots, n$, 都有

$$E(R_i) = \frac{n+1}{2}, \text{Var}(R_i) = \frac{n^2-1}{12}.$$

同样地, R_i 和 $R_j (i \neq j)$ 的联合分布也是均匀分布

$$P(R_i = r_i, R_j = r_j) = \frac{1}{n(n-1)},$$

其中 $r_i \neq r_j$, 从而有:

定理2 对任意的 $1 \leq i < j \leq n$, 都有 $\text{Cov}(R_i, R_j) = -\frac{n+1}{12}$.

在许多情况下, 数据中有相同的数字, 称为结 (tie). 结中数字的秩为它们按升幂排列后位置的平均值. 比如数据 2, 3, 3, 6, 10 这五个数的秩分别为 1, 2.5, 2.5, 4, 5. 也就是说, 处于第二和第三位置的两个 3 得到秩 $(2+3)/2=2.5$. 这样的秩称为中位秩. 如果结多了, 零分布的大样本公式就不准了, 因此需要修正.

附录 B R的图形界面

§B.1 R Commander

不同于S-PLUS, R自带的RGui没有提供专门的用于统计分析的菜单. 然而John Fox基于R开发了一套进行基础统计分析的菜单驱动的分析系统, 称为R Commander: A Basic-Statistics GUI for R. 有关的信息可参见John Fox的主页.

<http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/index.html>

下面简单介绍一下R Commander的安装、功能与使用等.

B.1.1 功能

R Commander是一个交互式菜单/对话框系统(menu/dialog-box interfaces), 用于进行数据的读、写、转换及常用的统计分析. 作者还添加了线性与广义线性模型等统计分析工具.

B.1.2 (网络)安装

R Commander的网络安装比较方便, 但需要较长的时间. 其步骤如下:

- 启动R. R缺省的安装为MDI模式, 建议改为SDI模式. 这可通过菜单“编辑”下的“GUI选项”设置;
- 点击菜单“程序包”=> “安装程序包”
- 选择一个较快的镜像站点(CRAN Mirror)
- 选择Rcmdr 安装. 期间会自动安装其它必要的程序包,时间较长!

B.1.3 运行

- 1) 方法1: 在RGui下通过“程序包”=>“载入程序包”加载Rcmdr程序包;
- 2) 方法2: 在RGui的命令窗口键入命令

```
> library(Rcmdr)
```

此后就激活R Commander, 如图B-1所示.

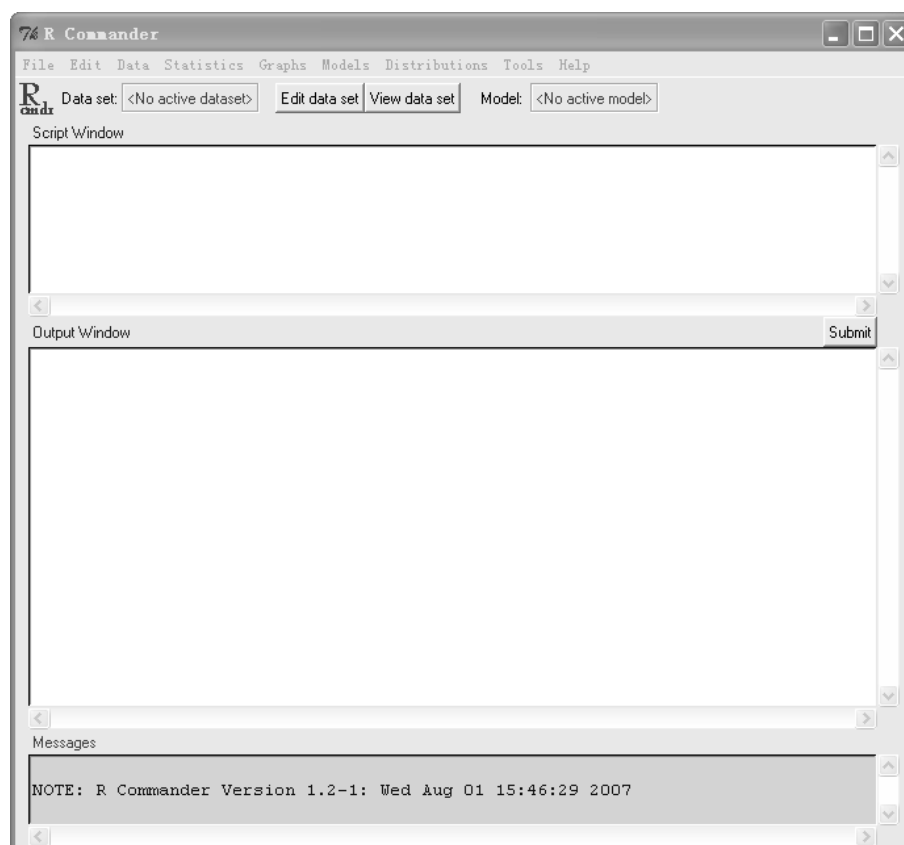


图 B-1 R Commander的窗口.

B.1.4 结构与使用

R Commander窗口从上到下的组成如下:

- 主菜单(Menu), 包括: File, Edit, Data, Statistics, Graphs, Models, Distribution, Tools 和Help.
- 工具条(Tool bar), 包括: Data set, Edit data set, View data set 和Model.
- 命令(代码)窗口(Script Window). 通过菜单所进行的操作的R代码在这里显示出来, 并立即被执行. 在这里你也可以修改已有的代码, 也可以输入自己的命令, 按Script窗口右下方的Submit按钮就可发送命令让R执行. 通过菜单进行的统计分析指向(激活)一个数据集. 一旦读入一个新的数据集, 它就被激活.
- 输出窗口(Output Window): Script窗口中执行的命令将在Output窗口中重新以红色显示出来, 并且给出相应的结果. 如果是作图, 则启动R Graphics页面.
- 信息窗口(Messages): 这里主要列出代码运行时出现的错误信息, 关以红色显示.

详细请阅读随R Commander安装的Help下的Introduction to R Commander: Getting Started With the R Commander(John Fox, 2006).

在R Commander 中进行数据分析的步骤如下:

- 1) 通过Data菜单建立或载入数据. 之后在Data set左侧出现数据集的名字;
- 2) 通过菜单的Statistics, Graphs, Models, Distributions等进行有针对性的分析. 菜单的有些项目是灰色的, 表示此项当前不可使用. 具体参看帮助文件的菜单树(Menu tree).

§B.2 PMG

下面主要就PMG的功能、安装与基本的使用方法作一简单的介绍.

B.2.1 功能

PMG为另一款用于基础统计分析的菜单驱动分析系统, 但与R Commander不同的是其动态的对话框, 即我们可以用鼠标的拖拉完成一系列的作图与具体的统计分析工作, 包括

\item 描述性统计分析
\item 数据的概括
\item 常规的统计推断
\item 线性回归分析

B.2.2 安装

PMG的安装与使用命令分别为:

```
> install.packages("pmg", dep=TRUE)
> require(pmg)
```

此后就激活PMG, 如图C-1所示.

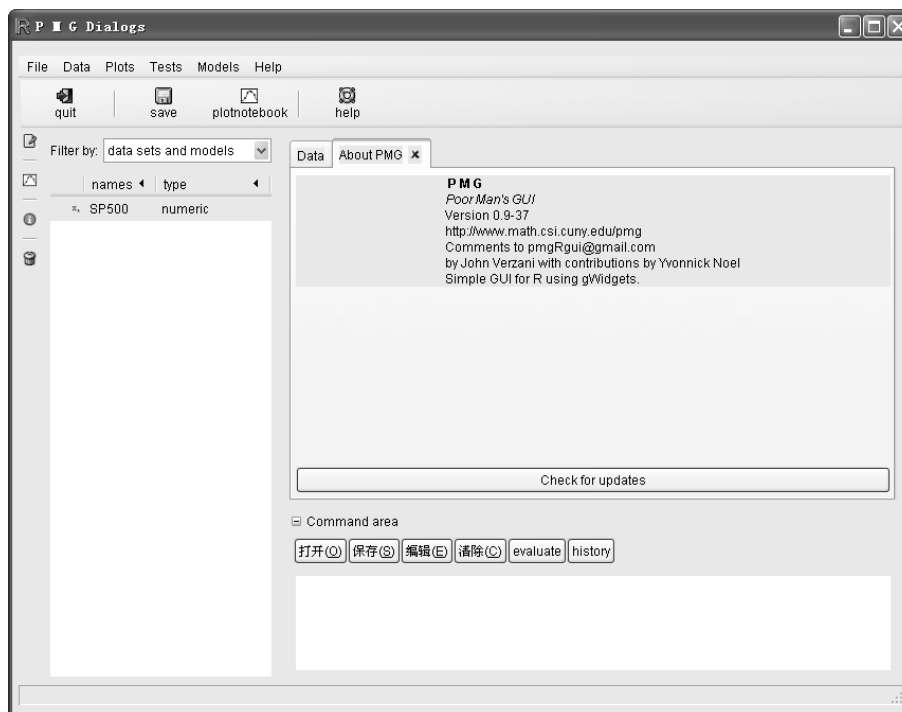


图 B-2 PMG的窗口.

B.2.3 结构与使用

PMG窗口从上到下的组成如下:

- 窗口上方第一排的主菜单(Menu): 包括File, Data, Plots, Tests, Models 和Help.
- 窗口上方第二排的工具条(Tool bar): 包括quit, save, plotnotebook 和help.
- 窗口左侧的快速拖动区域(quick drop area): 从上到下各按钮分别为数据的编辑(data.entry())、数据的作图(plot())、数据的概括(summary())和数据的移除(rm()).
- 下侧的命令区域(command area): 在这里你可以输入或复制自己的命令, 按evaluate按钮可发送命令让R执行. 上面通过拖拉的方式运行的命令和结果也在这里显示出来.
- 中间较大的为对话区域: 许多带选项的命令, 如boxplot()的选项会在这里出现, 等待你的给出(特出拉或直接输入).

下面举例来说明PMG的动态拖拉式操作过程.

- 1) 打开数据集: 通过“Data”⇒“Load data set...”打开数据集women;
- 2) 计算变量weight的均值: 通过“Data”⇒“Univariate summaries”⇒“mean”在对话区域打开函数mean()及选项; 将women的变量weight用鼠标拖拉到x=处, 再按“确定”按钮. 结果得到如图C-2所示的显示.
- 3) 计算并作出分位数: 通过“Data”⇒“Univariate summaries”⇒“quantiles”在对话区域打开函数quantile()及选项; 将women的变量weight用鼠标拖拉到x处, 再选择probs的一组值, 譬如probs: c(0.25, 0.5, 0.95), 结果得到如图C-3所示的画面.

其它功能与用法, 可通过菜单各条目的在线帮助或尝试了解, 在些不再细说.

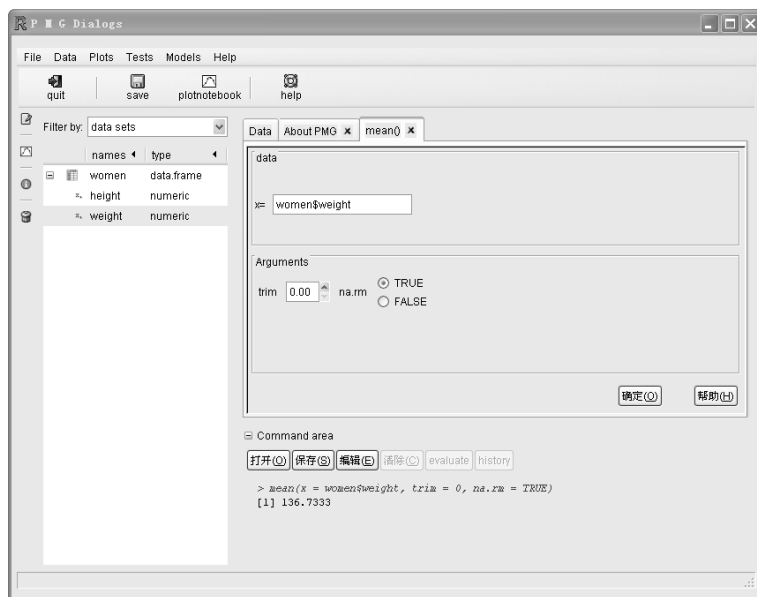


图 B-3 PMG求样本均值时的窗口.

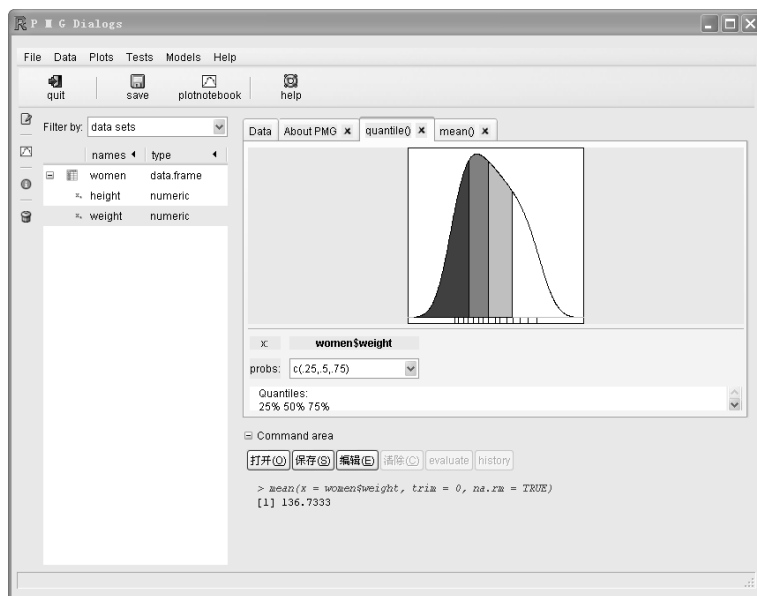


图 B-4 PMG求分位数时的窗口.

附录 C R的编程环境

§C.1 R WinEdt

优秀的程序编辑器很多, 这里我们介绍三款适合于R的编辑器, 供大家选择.

WinEdt 在是一款非常流行的源代码编辑器, 特别是它与 \LaTeX 组合后的CT \TeX 套装软件已为科技排版人员熟悉. WinEdit同样适合于R, Uwe Ligges开发的R WinEdt融合了WinEdt的优点, 并添加了R的菜单和工具条, 由此可以大大提供编程的效率. 下面对R WinEdt的安装、功能与使用等作一简单的介绍.

C.1.1 (网络)安装

R WinEdt 的网络安装比较方便, 但需要较长的时间. 其步骤如下:

- 在SDI模式启动R
- 点击菜单“程序包”=>“安装程序包”
- 选择一个较快的镜像站点(CRAN Mirror)
- 选择RWinEdt安装. 在这过程中请选择添加桌面快击.

C.1.2 运行

- 1) 方法1: 在RGui下通过“程序包”=>“载入程序包”加载RWinEdt程序包;
- 2) 方法2: 在RGui的命令窗口键入命令


```
> library("RWinEdt")
```

3) 方法3: 点击桌面的RWinEdt快击键.

此后就激活RWinEdt, 如图C-1所示. R WinEdt中的R菜单及工具条如图C-2所

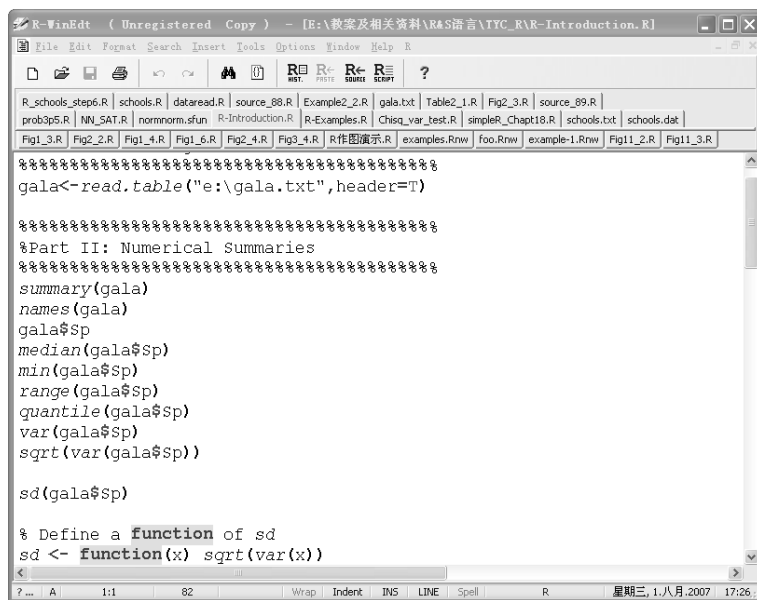


图 C-1 R WinEdit的窗口.

示.

C.1.3 R WinEdt的特点

- 与RGui共同运行
- 具有WinEdt的强大功能(如Delimiter检查, 高级搜索, 书签, 宏, 缩进与注释的对齐等)
- 语法高亮显示(Syntax-Highlighting)
- 同时可以编辑多个R程序
- 设置简单快速的按钮与快击键(见表C-1)
- 将窗口中的R文件(文件的所有代码)发送到R中运行

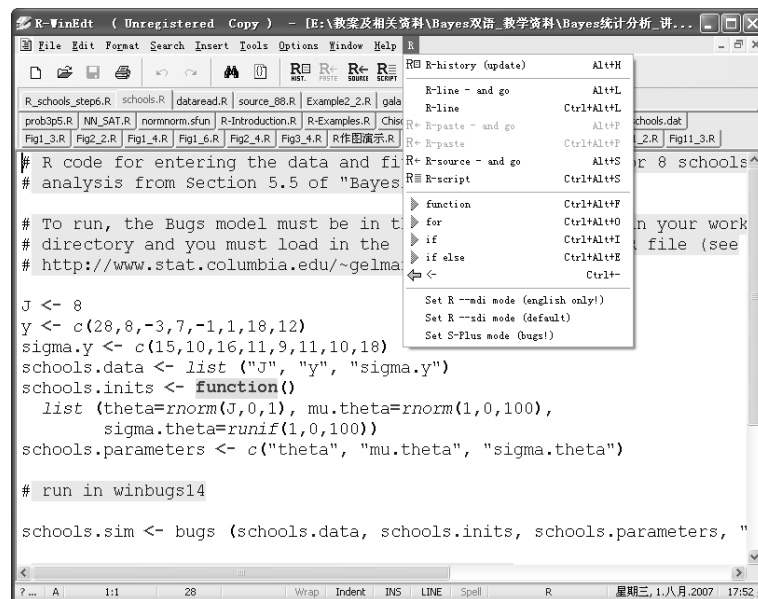


图 C-2 R WinEdit中R菜单.

- WinEdt中选中的代码发送到R中运行
- 单行代码发送到R中运行
- 及时更新历史命令记录文件.Rhistory, 以便重复使用旧的命令
- 提供结构化的模块, 如: `for(_ in _){_}`

C.1.4 R WinEdt的菜单与热键

§C.2 Tinn-R

Tinn也是一个很好的文本编辑器, 它与R的结合与R WinEdt类似.

- 1) 下载: <http://www.sciviews.org/Tinn-R/>
- 2) 安装: 点击Tinn-R x.xx.x.x setup.exe
- 3) 运行:
 - 启动R

表 C-1 R WinEdt的菜单与热键

命令	热键	菜单图标	说明
Brackets Check	Ctrl+F12		括号配对检查
R History	ALT+H	R HIST.	保存历史记录
R-line - and go	ALT+L		单行发送
R-line	Ctrl+ALT+L		单行发送并返回
R<- R-paste - and go	ALT+P	R<-PASTE	选中后发送
R<- R-paste	Ctrl+ALT+P		选中后发送并返回
R<- R-source - and go	ALT+S	R<-SOURCE	R文件发送(先打开)
R<- R-script	Ctrl+ALT+S	R SCRIPT	R文件发送并返回
function	Ctrl+Alt+F		生成函数框架
for	Ctrl+Alt+O		生成for循环框架
if	Ctrl+Alt+I		生成if框架
ifelse	Ctrl+Alt+E		生成ifelse框架
<-	Ctrl+-		生成赋值符号

- 点击快击按钮, 启动Tinn-R

4) 使用: 与R WinEdt类似. 与R WinEdt不同的一个特点是, Tinn-R的R card对于熟悉R中的函数及编辑是非常有用的. 更多关于Tinn-R, 见Tinn-R FAQ:

http://www.sciviews.org/Tinn-R/Tinn-R_FAQ.html

Tinn-R中的R菜单及如图C-3所示.

§C.3 SciViews R

SciViews R与Tinn-R类似, 它是由Philippe Grosjean, Eric Lecoutre, JoséCláudio Faria, Marta Rufino开发.

- 1) 下载: <http://www.sciviews.org/SciViews-R/>
- 2) 安装: 点击SciViews-R_x.x-xx Setup.exe
- 3) 运行: 点击快击按钮SciViews R Console

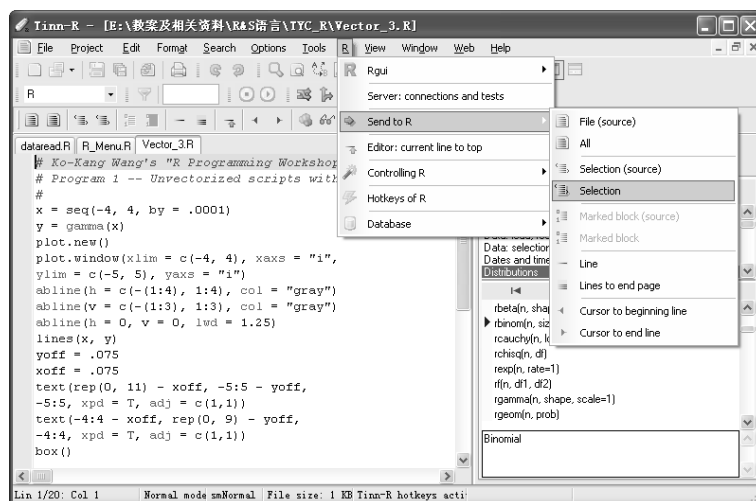


图 C-3 Tinn-R中R菜单.

注: 不同于R WinEdt和Tinn-R, 启动SciViews R 会自动启动R, 并使之成为SciViews R的一部分.

- 4) 使用: 点击文本编辑器左侧的小三角形就可执行光标所在的行或选中的代码段. 详见SciViews R主页的手册.

SciViews R中的窗口布局如图C-4所示.

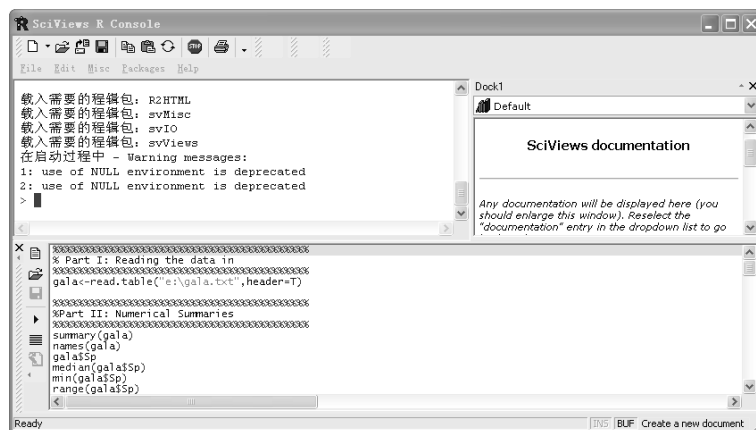


图 C-4 SciViews R的窗口布局.

参考文献

- [1] Peter Dalgaard, Introductory Statistics with R, Springer, 2002.
- [2] John Maindonald, John Braun, Data Analysis and Graphics Using R — An Example-based Approach, Cambridge University Press, 2003.
- [3] John Maindonald, Using R for Data Analysis and Graphics — Introduction, Examples and Commentary, 2004.
(<http://www.ats.ucla.edu/stat/R/sk/booksusingr.htm>)
- [4] John Fox, An R and S-Plus Companion to Applied Regression, Sage Publications, Inc., 2002.
- [5] Julian J. Faraway, Linear Models With R, Chapman & Hall/CRC, 2004.
- [6] Julian J. Faraway, Extending the Linear Model with R, Chapman & Hall/CRC, 2006.
- [7] Julian J. Faraway, Practical Regression and ANOVA Using R, 2002.
<http://www.stat.lsa.umich.edu/faraway/>,
<http://www.ats.ucla.edu/stat/r/sk/bookspra.htm>
- [8] John Verzani, Simple R — Using R for Introductory Statistics, 2002.
(<http://www.math.csi.cuny.edu/Statistics/R/simpleR/index.html>)
- [9] W. N. Venables, D. M. Smith and the R Development Core Team, An Introduction to R — Notes on R: A Programming Environment for Data Analysis and Graphics, 2006. (中译本1.7—R语言介绍, 中译本2.3.0—R导论)
- [10] The R Development Core Team, R Reference Manual. (R: A Language and Environment for Statistical Computing — Reference Index), 2006
- [11] The R Development Core Team, R Data Import/Export, 2006. (中译本2.2.1—R数据导入与导出)

-
- [12] Emmanuel Paradis, R for Beginners, 2005. (中译本2.2.1, 2006)
- [14] Søren Højsgaard and Ulrich Halekoh, R in a few hours - a brief introduction, 2006.
- [14] Ulrich Halekoh, Jørgen Vinsløv Hansen, Søren Højsgaard, Basic graphics in R, 2006.
- [15] <http://cran.r-project.org/doc/FAQ/R-FAQ.html>.
- [16] Ihaka R & Gentleman R.R: A Language for data analysis and graphics [J]. J Comput Graph Stat, 1996; 5(3): 299-314.
- [17] Eric Zivot and Jiahui Wang, Modelling Financial Time Series with S-PLUS, 2002.
- [18] Paul Murrell, R Graphics, Chapman & Hall/CRC, 2006.
- [19] Jim Albert, Bayesian Computation with R, Springer, 2007.
- [20] A. Gelman, J.B. Carlin, H.S. Stern and D.B. Rubin, Bayesian Data Analysis, Chapman & Hall/CRC, 2004.
- [21] S-Plus应用统计教程, 上海财经大学出版社, 2005.
- [22] 茆诗松, 周纪芃, 概率论与数理统计, 中国统计出版社, 1999.
- [23] 茆诗松, 程依明, 濮晓龙, 概率论与数理统计教程, 高等教育出版社, 2004.
- [24] 何书元, 概率论与数理统计, 高等教育出版社, 2006.
- [25] 王岩, 隋思涟, 王爱青, 数理统计与MATLAB工程数据分析, 清华大学出版社, 2006.
- [26] 梁小筠, 祝大平, 抽样调查的方法和原理, 华东师范大学出版社, 1994.
- [27] 梁之舜, 邓集贤, 杨维权, 司徒荣, 邓永录, 概率论与数理统计(下), 高等教育出版社, 2003.
- [28] 王静龙, 梁小筠, 非参数统计, 高等教育出版社, 2006.
- [29] 王静龙, 梁小筠, 定性数据分析, 华东师范大学出版社, 2005.
- [30] 吴喜之, 非参数统计, 中国统计出版社, 1999.
- [31] 吴喜之, 统计学: 从数据到结论, 第二版, 中国统计出版社, 2004.
- [32] 王星, 非参数统计, 中国人民大学出版社, 2005.
- [33] 高惠璇, 实用统计方法与SAS系统, 北京大学出版社, 2001.
- [34] 方开泰, 实用多元统计分析, 华东师范大学出版社, 1989.
- [35] 薛毅, 陈立萍, 统计建模与R软件, 清华大学出版社, 2007.
- [36] 陈毅恒, 梁沛霖, R软件操作入门, 中国统计出版社, 2006.
- [37] 耿修林, 应用统计学: 学习指导、软件介绍及习题, 科学出版社, 2004.

本书特色

统计学以数据为研究对象,它是以概率统计为基础、应用统计学的基本原理和方法并结合统计软件对实际数据进行收集、整理和分析的一门科学.数据的统计分析涉及大量的统计计算,包括向量与矩阵的运算,这时传统的手工或基于计算器的计算几乎无法进行,必须借助于现代化的计算工具—统计软件,特别是在参数不断增多、数据维数不断增大、变量之间相关性不断密切的经济、金融、生物、制药、社会、心理等领域,统计软件的使用显得尤为重要.因此我们不仅要通过数理统计这门课程的学习掌握统计学中的基本理论与方法,更应该将这些理论与方法运用于实践,并通过统计软件用图形直观展示数据中存在的特征,用具体的统计方法揭示其中存在的规律,解决一些具体的实际问题.作为自由、免费、源代码开放、维护更新及时的软件、其强大的图形展示和统计分析功能,使R语言成为学好数理统计最好的工具.

作为数据统计分析的教科书,本书有如下几个特点:

- 1) R软件介绍精简实用,自成一体;
- 2) 原理讲解与软件使用高度结合;
- 3) 内容全面,涵盖统计各学科需要的主要统计方法;
- 4) 内容安排循序渐进,又相对独立,不失为数据统计分析的工具书;
- 5) 突出对原理与方法的理解、更注重实例通过R的求解过程和对结果的解释.
- 6) 全书使用L^AT_EX编辑排版,印刷质量一流,是中英文L^AT_EX排版的经典作品.